

合理性的本质

〔美〕罗伯特·诺奇克 著　葛四友　陈昉 译

二十世纪西方哲学经典

Robert Nozick

The Nature of Rationality

海译文出版社

世纪出版

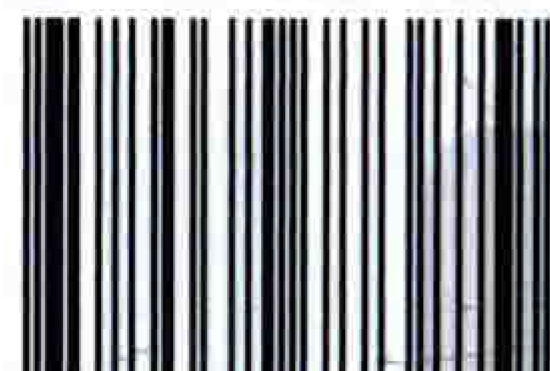
《二十世纪西方哲学经典》选收二十世纪西方哲学界各主要流派影响较大的著作，通过有选择的译介，旨在增进文化积累，拓展学术视野，丰富研究课题，为了解和研讨现代西方哲学提供系统而完整的第一手资料，以利于理论界、学术界深化对西方文化的研究和借鉴。

《合理性的本质》围绕当代哲学、社会科学诸多领域争论的核心概念——合理性来展开讨论，全面探究了当代西方学术界从进化论视角对合理性的理解，特别是从合理性角度来思考原则的选择问题，又依据原则来思考合理性的问题，以决策价值将两者结合更是让人眼前一亮。

上架建议：外国哲学

ISBN 978-7-5327-7387-9





定价：98.00 元

上海译文出版社
www.yiwen.com.cn

易文网：www.ewen.co



合理性的本质

〔美〕罗伯特·诺奇克 著　葛四友　陈昉 译

二十世纪西方哲学经典

Robert Nozick

The Nature of Rationality

上海译文出版社

图书在版编目(CIP)数据

合理性的本质/(美)罗伯特·诺奇克(Robert Nozick)著;
葛四友,陈昉译. —上海:上海译文出版社,2016.12
(二十世纪西方哲学经典)
书名原文:The Nature of Rationality
ISBN 978-7-5327-7387-9

Ⅰ.①合… Ⅱ.①罗… ②葛… ③陈… Ⅲ.①理性—
研究 Ⅳ.①B017

中国版本图书馆 CIP 数据核字(2016)第 226919 号

Robert Nozick
THE NATURE OF RATIONALITY
Copyright © 1993 by Princeton University Press
All Rights Reserved.
No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage and retrieval system, without permission in writing from the Publisher.
图字:09-2005-551 号

合理性的本质
[美]罗伯特·诺奇克 著 葛四友 陈昉 译
责任编辑/张吉人 装帧设计/张志全工作室

上海世纪出版股份有限公司
译文出版社出版
网址:www.yiwen.com.cn
上海世纪出版股份有限公司发行中心发行
200001 上海福建中路 193 号 www.ewen.co
江阴金马印刷有限公司印刷

开本 890×1240 1/32 印张 10.25 插页 5 字数 248,000
2016 年 12 月第 1 版 2016 年 12 月第 1 次印刷
印数:0,001—1,000 册

ISBN 978-7-5327-7387-9/B·433
定价:98.00 元

本书中文简体字专有出版权归本社独家所有,非经本社同意不得转载、摘编或复制
本书如有质量问题,请与承印厂质量科联系。T:0510-86683980

献给 Carl Hempel

并以此纪念 Gregory Vlastos

致　　谢

本书的前两章是1991年12月13日和15日在普林斯顿大学所做的唐纳讲座。我是普林斯顿大学毕业的，正如这本书一样，这两个讲座也是献给我的老师们的。第1、2章的重刊得到了犹他大学的许可，选自 *Tanner Lectures on Human Values*, vol. 14 (Salt Lake City: University of Utah Press, ©1992)(这里刊出的版本做了一点增补和改变)。这两章的第一稿是1989年夏天在意大利的 Bellagio 的洛克菲勒基金研究中心所写的。

第5章的部分内容构成了1990年3月在 Pacific Lutheran University 所做的 the Walter C. Schanckenberg Memorial Lecture。第3—5章的部分内容是1992年5月在 the University of Chicago 所做的 Centennial Lecture。

我很感激在普林斯顿大学的各位讨论者所提出的评论和建议——Gilbert Harman（他还阅读了整个手稿），Clifford Geertz, Susan Hurley, 和 Amos Tversky——也感谢 Scott Brewer, Eugene Goodheart, David Gordon, Christine Korsgaard, Elijah Millgram, Bill Puka, Tim Scanlon, Howard Sobel 和 William Talbott。特别要感谢 Amartya Sen 他和我一起开了课，我们在课内课外做了很多有激发性的讨论。

我非常感激 Laurance Rockfeller 对此研究项目有兴趣，给予慷慨支持。

我谢谢我的妻子，Gjertrud Schnackenberg，她使得写此书的这些年变得如此地浪漫、可爱和有趣。

目　　录

导　论[1]

哲学一词意指爱智(wisdom),但是哲学家们真正爱的却是推理(reasoning)。他们构建理论,并组织各种理由加以佐证;他们考虑各种反对意见,并努力回应它们;他们还阐发各种反驳其他观点的论辩。即使是那些宣称理性(reason)有限的哲学家(如古希腊的怀疑论者们、大卫·休谟以及质疑科学之客观性的论者们),也无一不是提出各种理由来支撑自己的观点,并提出相反观点的各种棘手问题。而宣言或者格言警句却并不被视作哲学,除非它们也崇尚并勾画推理。

哲学家推理的对象之一就是推理本身。推理应当遵循什么样的原则?它必须遵循什么样的原则?亚里士多德最早对演绎原则进行了明确的阐释和研究;科学和概率论的研究者勾画了各种非演绎的推理和支持模式;笛卡儿试图表明为什么我们应当信任推理的结论,而休谟则质疑我们这样做的合理性(rationality);康德也勘定了他所认为的理性的恰当领域。过去,对于理性的这种勾画并非一项学术工作。各种新发现的观点都将得到应用,因为人们的推理将会得到改进,他们的信念、实践和行为也会变得更合理。苏格拉底发现,对当下的信念与

---

① 本部分为邓正来先生与陈昉博士合译,葛四友为了全书主要术语翻译的前后一致,做了稍许变动。——译者

实践的合理性进行探究，有着很多的风险。一个社会的各种传统有时候是经不起认真检视的，而且也并不是所有的人都愿意看到“显白地”考察“隐晦物”（the implicit）。即便是简单地考虑一下其他可能的选择，似乎也会成为一种对现状具有侵蚀力的颠覆，亦即一种对专断的揭露。

古希腊人认为，理性确定了人类的独特性。“人是一种理性的动物。”理性能力把人与其他动物区分了开来，并由此而定义了人。然而，人类的这一特性自中世纪以降不断地被缩小——这是我回想起读到过的最早的关于智识（intellectual）史的宏大论述。哥白尼、达尔文和弗洛伊德都教导我们说，人类在宇宙中并不占据某种特殊的地位，人类在起源上也不具有特殊性，而且其行为也并非总是受理性动机的指导，甚或并非总是受那些在意识上可知动机的指导。尽管如此，持续赋予人类以某种特殊地位的依然是人类的理性能力。也许，我们并不是一以贯之地践履这一可贵品性，但正是它使得我们具有了独特性。理性为我们提供了去探究和发现每一样东西的（潜在）能力；它也使得我们能够经由理由和援用原则来控制和指导我们的行为。

xii 因此，理性乃是人类自我形象的一个关键成分，而不仅仅是我们获取知识或改善我们生活和社会的一项工具。对我们理性的理解会使我们更为深刻地洞见到人的本性以及我们所拥有的那种特殊地位。希腊人把理性看作是独立于动物性的，因此肯定不是动物性的自然结果（outgrowth）。然而，进化论却使得有可能把理性看作诸种动物性特征之一，亦即一种具有有限目的和功能的进化适应性。

我相信，这一视角对哲学能够产生重大的影响。理性从来就不只是哲学家的专好，也不只是他们研究对象的一个重要部分；它始终还是哲学家用以发现真理（truth）的一项特殊工具，

亦即一种潜力无限的工具。[在《纯粹理性批判》一书中,康德赋予了理性一种相对卑微的作用:理性并不是要去认知一个独立实在(independent reality)的实质,而是要去认识一个由理性部分地构成和塑造的经验领域(empirical realm)。虽说如此,理性发挥作用的有效范围依然是极其宽泛的。]如果理性只是一种具有有限目的和功能的进化适应性,旨在与被理性视为当然且以之为基础的稳定事实相配合而发挥作用,而如果哲学又企图无限地适用理性(reason)且合理地(rationally)证成每一种信念和假设,那么我们也就可以理解,为什么许多哲学上的传统难题显得如此棘手且无法合理地解决。这些问题也许正是企图把合理性的作用扩展到其有限的进化功能之外才导致的。我在此想到的问题是指归纳问题、他心(other minds)问题、外部世界问题以及证成目标的问题。我将在后文中考察这种进化视角所具有的含义和后果。

近年来,合理性一直是受到特殊批判的一个对象。有人提出这样的主张,合理性是*有偏见的*,因为它是一个基于阶级的、男性的、西方或无论什么东西的观念。然而,致力于关注偏见(包括其自身的偏见)并力图控制与纠正这些偏见,正是合理性的组成部分。(也许试图纠正偏见这一企图本身就可能是一种偏见吗?但如果这是一种*批评*,那么它是从哪儿冒出来的呢?是否存在这样一种观点,即偏见是坏的,但是纠正这一偏见也是坏的?如果人们认为根除偏见是不可能的,那么偏见的指控又在何种意义上构成了一种批判呢?再者,这种不可能性所意指的是存在着某种在本质上无法被根除的特殊偏见,还是只是不能同时根除掉所有偏见呢?)

指控现有的标准中含有某种偏见,并不表明偏见就存在。因为这一结论,即现有标准(在一些被适用的情形中)自身表明

了某些特定的具体扭曲和偏见，乃是通过运用推理和证据——因而也就是在使用我们现有的标准——而达致的。因此，仅仅说我们（所有的人）都是透过我们的概念体系来看世界的，乃是不充分的。问题在于：我们特定的概念体系与标准究竟是以何种特定方式和通过何种确切机制而造成扭曲的？而一旦有人向我们表明了这一点，我们就可以着手纠正了。当然，我们现有的关于合理性的标准并不完美——而我们能指望这些标准在何年何月能如此呢？但是，这些标准有着真正的优点，因此如果有人想表明它们是有缺陷的，那么他就至少需要拿出与这些正遭受抨击的标准具有同等分量的合理论证才行。发现这类特定的缺陷，乃是通往纠正它们并更恰当地构建这些合理性标准的必要的第一步。因此，应当欢迎且尽力探寻指控合理性标准之中存有偏见的证据。合理性标准乃是我们据以超越或者审查我们自己特有的希求、愿望和偏见的一种手段。如果当下广为流行的对合理性标准的批判所导致的结果乃是否弃或颠覆人类据以能够纠正并超越个人及群体之偏见的一个主要手段的话，那么这将既可笑又可悲。

关于合理性的研究——它无论是对于个人还是对于社会来讲都具有极其重大的评价意义和实践意义——如今已然被转变成了一个技术性主题。原则要更加明确以勾画有效的推理和把握各种有理据支撑的信念与行为的模式。演绎逻辑在 19 世纪晚期被哥特洛布·弗雷格(Gottlob Frege)所改造，并在 20 世纪迸发为一种技术表述。我们在过去发展出了逻辑体系，而且使用逻辑技巧来探究逻辑体系自身的属性与局限。概率论导致了统计学推理的形式理论，且在把信念合理性理论化和构建归纳逻辑（至少是若干有关接受的归纳规则）的基本原理的尝试中，数学无处不在。在本世纪，数学家、经济学家、统计学家和哲学

家发展了一种精密而强有力的合理行为理论——亦即决策理论,而现在,这一理论已被广泛运用于各种各样的理论语境与实践语境中。(这一理论装置为合理策略互动的形式理论、博弈论、社会选择与福利经济学的形式理论、微观经济现象的理论以及有关政治领域的各种系统理论提供了框架。)在相关文献中充满了——如果说还没有被完全吞没的话——在各种数学结构中以陌生的象征符号所组成的难解公式。我并不是在诋毁这种转向。当今这些理论上的发展乃是与此前的理论动机和关注一脉相承的,而且还大大推进了此前的研究。

本书也将考虑这些术语(technicalities),并打算从合理性理论所覆盖的两大领域[即"决策的合理性"(rationality of decision)和"信念的合理性"(rationality of belief)]中提出一些新的术语。我们拟重构现有的决策理论,使之能够含括行动的象征意义;我们拟提出一项新的合理决策规则[决策价值(decision-value)最大化规则];然后,我们还将着手探析这项新规则对于"囚徒困境"与"纽科姆难题"(Newcomb's Problem)所具有的意义。信念合理性涉及两个方面的内容:第一,得到那些使信念可靠的理由的支撑;第二,由一种能可靠地产生出真信念的过程而生成。(我提出的用以解释这两个方面之间所具有的那种令人困惑的关联的进化论说逆转了康德的"哥白尼式革命"的方向。)我将提出两项规则来管辖"合理的信念":不相信可信度低于不相容替代项的任何一项陈述——智识成分;只有在相信该陈述比不相信该陈述所能达到的预期效用(或决策价值)更大时才相信——实践成分。于是,这一双重结构就被应用于各种有关"信念伦理"的问题,且对"摸彩悖论"(lottery paradox)问题提供一种新解。此外,我还将考察工具合理性(亦即对给定目标的有效且有效率的追求)的范围及局限,并给目的

xiv

合理性提出一些新的条件。因为合理思考还包括构建出各种新的且富有启示意义的哲学问题和思想,所以这里也会提出这样做的一些启发法(heuristics)。故而,本书将充斥着为进一步推进合理性的根本问题上的思考而有必要的那些技术细节。

然而,人们仍有某种理由担忧。在此之前,关于合理性的各种问题一直都是人类的共有话题,有时候还以颇为繁复的思维方式加以探讨——没人会宣称康德的《纯粹理性批判》是一本易懂的书——尽管如此,只要愿付出努力的话,知识分子在很大程度上还是能够理解这些问题的。过去,研究这些问题的各种新思想都会变成大众文化的组成部分;它们塑造了讨论与争辩的术语,有时候甚至塑造了感知术语[回忆一下康德的思想对柯尔律治(Coleridge)产生了多么大的影响]。但是现在情形不同了——而且不限于合理性这个论题。

对有关人类根本关注的许多论题而言,最有成效和最有益的探讨路线是越来越有技术化的转向了。今天,如果我们不把握这些技术性发展,不掌握这些发展所开放出来的新问题,不了解某些传统立场被颠覆的方式,那么我们就不可能充分地探讨这些论题。不列颠百科全书最近出版了《西方世界巨著丛书》(*Great Books of the Western World*)第二版,然而此举却在下述两个问题上引发了公开的争议:一是关于女性及少数群体论题的表达(或者说是相对缺乏)的问题;二是关于任何一部"巨xv 著"被公认为是精英产品的问题。[①] 然而,对于许多本世纪最伟

① 我本人认为,将许多不同论者的论著统一出版,并用一个比其中任何一本著作的书名或任何一位著者的姓名都更卓越醒目的丛书名头来加以颂扬的做法,并不是对该人已完成的思想成就的一种妥适表达。但是,如果某一团体去出版此类书籍的一个书目并重印那些不易找到的书籍,那仍可能是一种有助益的努力;当然,不同的团体可以出版不同的书目。

大的智识著作未予收录这一事实,却无人置评,原因大概是这些著作对于那些只受过一般教育的读者来说太过技术化了。

关键并不仅仅在于本世纪所产生的值得人们关注的思想和研究成果无法被即便受过良好教育之人群中的大多数人所理解——因为自牛顿以来实际情况一直就是这样的,而毋宁是当今的这些思想所关注的乃是我们想理解和需要理解的那些论题,亦即我们认为每一个人都应当理解的那些论题。但是,如果我们不对这些术语有所了解的话,那么我们就无法理解或合理地探讨这些论题。我们的评价术语本身就已经变得技术化了。

下面我将列举一些业已经历技术化发展的论题:第一,公共福利的观念[和卢梭的"公意"(general will)观念]以及对于民主投票程序之目的的理解,都被"阿罗不可能定理"(Kenneth Arrow's Impossibility Theorem)所转变了。这个定理向我们表明:若干极为自然且可欲的条件,显然应当为任何用来决定公共福利或民主地最偏好选项的程序所遵守,但它们却是无法被同时满足的。因此,有些条件必须被放弃掉。第二,阿玛蒂亚·森(Amartya Sen)关于"帕累托自由悖论"(Paretian liberal paradox)的研究表明:一种非常自然的关于个人权利和个人自由权项之范围的解释,与一种同样自然的关于各种社会选择应当如何以合理的方式组织起来的解释,这两者并不容易和谐共处。因此,这些观念需要一种新的结构化。第三,物理世界的基本性质——时空结构——不可能脱离开广义相对论所提出的关于时空的技术性(和数学)而得到理解。第四,就因果关系的性质以及物理世界的独立特性而言,由于它们乃是由我们目前所拥有的最精确且最成功的科学理论(即量子场论)所描画,所以情形亦复如此。第五,对数学真理——自古希腊以来即是我们最好与最确定的知识之典范——之性质与地位的探讨已然在极

大的程度上被哥德尔(Kurt Godel)的“不完备定理”所改变了。第六,关于“无限性”(infinity)及其各个层次的性质现在已在当代的集合论(set theory)中得到了阐述和探究。第七,如果没有关于一种价格机制及与之相配套的各种私有产权制度是如何使合理的经济计算成为可能的理论,又如果没有持续数十年的关于在一个社会主义社会里究竟是否可能进行合理计算的学术争论,我们就无法理解为什么那些共产主义社会的经济效率会如此之低。第八,在个人合理性以及人与人之间合理互动的方面,也已经出现了许多理论上的进展:决策理论、博弈论、概率论和各种统计推理的理论。

在上述任何一个论题中,20世纪都已经贡献出了种种崭新的研究成果和理论,而如果人们缺乏对某些技术结构和细节的理解,那么这些东西将难以理解或者难以可靠地讨论。我意识到,这只是一个哲学家的问题清单;而社会科学家和自然科学家则会在这个单子上添加更多的论题。这一点增强了我的观点。知识分子、受过教育的人和严肃的人所具有的那种一般文化已经不再能够把握许多论题了,而这些论题对于理解和思考社会、人类以及整个宇宙来说都是至关重要的。耳熟能详的是,有许多复杂的科学事实性问题是必须求助于那些可能会意见不一的(比如说在有关各种实践活动的环境影响上)专家的。新颖的东西在于:我们希望用以进行评价和理解的许多术语及概念本身已经变得技术化了。

我提出了这个问题,但却没有给出解决方法。当然,这些材料的展现对于一般读者来说仍是有必要的。但是,对这些材料最为清晰的展现方式——如果它确实是要准确地传达那些基本思想的话——将含括一些专门性的描述与推进,而这因此也会限制读者的范围。这个方面的要求对于一部以呈现并探究新观

点为目的的著作来说就更加困难了。我不想让这个关于合理性的论题远离一般读者的视野。然而,有些观点只能以一种或多或少技术化的方式予以陈述、规定和辩护。我已经努力把这些技术化细节减至最低,或至少已经努力把它们限制在特定的章节中了。为了我们这个社会的智识健康——更别提我们这些知识分子的社会健康了,那些基本的思想还是必须保留在公共的视域当中。

1. 如何以原则做事

原则是为了什么？为什么我们要持有原则？为什么我们要提出它们？为什么我们又要遵守它们呢？我们本可以只凭突发的奇想或者一时的激情而做事。我们也可以为了最大化自己的个人利益而做事，并且建议其他人也这样做。那么原则究竟是对心血来潮和私利的一种约束，还是遵循原则不过是增进自我利益的一种手段呢？原则究竟具有什么样的功能呢？

行动原则把行动进行归类，把它们归置于一般的类目之下；然后由此关联的行为就要得到相同方式的对待或者处理。这种一般性（generality）能服务于几个不同的功能：智识的（intellectual）、人际间的（interpersonal）、内省的（intrapersonal）和个人的（personal）。下面我将从智识方面的功能讲起。

智 识 功 能

我们可考虑司法判决。在一种设想的体系中，法官的断案是简单的，就是在具体案件中产生她认为最好或最可取的结果。另一种司法判决体系涉及的是有原则的判决：判例法（common law）的法官首先会构建出（formulate）符合（大部分或绝大部分）先例和一系列虚拟案件的原则，然后再运用这一原则来判决

目前的案件。[1] 这种试图构建可接受的一般性原则,就是检验(test)对具体案件所做的判决:是否存在某个适当的一般性原则,它在已知案件和一目了然的虚拟案件中都能得出正确的结果,且它在目前这一案件中也能产生你想要的那种结果。如果找不到这样的原则,那么就请重新考虑你在这个案件中想要何种结果。

这种程序对具体判断所做的检验乃是基于这样的假设,即任何一个正确的判断[2]都是由某个真的、可接受的一般性原则所产生的,也就是说,各个真的具体判断乃是将一般性原则应用于各种具体情境的结果。如果我们找不到产生某个具体判断的可接受的一般性原则,那这可能意味着不存在任何一个这样的一般性原则。该具体判断在此情形下就是错误的,并且应该放弃。不过这也许只是因为你不够聪明才未能构建出那项正确的原则。我们并没有机械的程序来决定何种解释才是正确的。[3]

---

① 我在此的目的不是要提出法律制度起作用的完整画面,而是通过与司法判决的某些方面做类比,强调原则在法律领域之外所具有的一般性特征。一个当下的司法判决是如何由一个契合于过去先例的原则所产生的,(在法律之外的)原则是如何产生一个正确的判断的,这两者之间的类比是有阐释力的。在法律体系之内,遵循先例(*stare decisis*)本身就是法律的(高阶)原则,它也可能与其他的原则相冲突或相竞争,但这不是我们现在要关注的。

② 当 judgment 和 case 用在法律语下时分别译为判决和案件,而用在一般语境下时译为判断和情形。对于 rationality 和 rational 的译法,基本的策略是,当它们直接用在人身上或描述人的属性时,译为理性或理性的,其他时候尽量译为合理性、合理或合理的。尽管诺齐克并没有特别区分 rationality 和 reason,尤其在引用时,两者更是交替使用。但这里还是用了合理性(rationality)与理性(reason)两个译法来区分。——译者。

③ 一个较弱的假设会认为,并不是每一个正确的判断都是由一个可接受的原则所产生的,但是有一些或者大部分是。然而,找出一个可接受的一般性原则,它能产生一个具体判断,这将(倾向于)表明该判断是正确的。尽管如此,找不到一项一般性原则也并不是抛弃该判断的结论性理由,因为它有可能属于那种孤立存在的判断,并非任何可接受的原则的结果。

若你能找到包含这一情形的一般性原则或理论，愿意将此原则适用于其他情形，那么这个具体判断就获得了新的支持。请考虑经验数据点 a、b、c 和 d，如果贯穿它们的最简单的线是一条直线的话，那么这将支持另一点 e 也在这条直线上的预测。归纳逻辑学者们已经发现，要隔离且解释（相对）简单的一个法则式（lawlike）陈述对现存的数据点如何归类，以至于能对新的数据点做出合法的推理和预测并不容易。尽管如此，我们并不怀疑，数据能够支持“一个法则是成立的”这一假说，且支持“一个新点会符合该法则”这一预测。同样，涵盖了 a、b、c 和 d 四个可接受的规范观点的最简单的原则，也将支持把另外一个（符合此原则的）判断 e 作为一个正确的规范观点。若理论家能够构建出与其具体判断相符合的一般性原则或理论，尤其是这个原则表面上也很有吸引力，那么他就会对自己做出的判断（或者在一个争议中所持的立场）充满信心。①

科学哲学家试图把科学法则与偶然的概括区别开来。偶然的概括仅仅是碰巧为真或碰巧一直为真。例如，从我的口袋里的硬币都是角币这样一种偶然概括出发，并不能推出这个假设陈述：*如果*我的口袋里又有了一枚硬币，那么这枚硬币一定*也*

---

① Mark Tushnet 主张，在法律领域，“判决要是有原则的”这一要求对法官所能达到的结果并不构成任何约束；如果先前的案件契合于一个原则（即使是一个已有的原则），但其结论是法官在当前案件中想避免的话，那么当前案件总是可通过这样或者那样的特征与其他案件区分开来。参见：Tushnet，“Following the Rules Laid Down：A Critique of Interpretavism and Neutral Principles”，*Harvard Law Review* 96（1983）：781－827。然而，仅仅将当前案件区别出来（充其量）只是允许了一个新判决，它并没有*支持*它。要支持它，法官就必须再构建出一个新的原则，且表面上是讲得通的，它要契合于（绝大部分）既往的案例、这个新案件和一些一目了然的虚拟案件。也就是说，对她做出的这一个区分，还有为什么她这个区分能够产生影响，她需要构建出一个有原则的理据。构建出一个可接受的原则并不是一件简单的事情，更不用说只要个人对新案件有这个想法，她就能频繁地这样做。

是角币。(另一方面,若从一个科学法则出发,比如说所有自由下落的物体运行的距离都等于 $1/2gt^2$,那么我们就可以推出,如果现在静止的某个物体自由下落 t 秒,其运行距离也会等于 $1/2gt^2$。)若先前的所有数据都契合于一项既定概括,那也仅当此概括的形式是法则式的,由此是一个法则的候选时,我们才能可行地(plausibly)推出:新的数据也将契合于该概括(且由此预测,收集到的那些新数据也将契合于它)。只有当数据落入法则式的陈述之下(或出现在几个法则式陈述之下)时,我们才可以合法地(legitimately)将其外推到进一步的情形。法则式陈述具有使其自身区别于偶然概括的一些方面,正是这些方面构成了一个通行证,允许我们从既定的数据来预测或预期进一步的数据。同样,对于具体的规范判断而言,核准我们基于先前判断而过渡到进一步判断的,正是这些既有判断都落在一般性规范原则之下。一项规范原则的诸特征核准我们对新的情形做出虚拟推断,这种新的情形超出了那些碰巧已经落在该原则之下的范例。原则是传递概率或支持的装置。从各种数据或情形得出的概率或支持,经由原则,传递到对新观察或新情形的判断和预测,要不然这些判断或预测就是我们不知道的,或不那么确定的。

然而,什么特征使得原则能够传递这种概率呢?在区分科学的法则式陈述[或通常的全称命题(nomic universals)]与偶然的概括时,我们会提到下述特征。[1] 法则式陈述中不包含具体的个体对象、日期、时段等术语;若包含的话,这些表述也能从

---

① 参见 C. G. Hempel: *Aspects of Scientific Explanation* (New York: Free Press, 1965), pp. 264 - 272;以及 Ernest Nagel: *The Structure of Science* (New York: Harcourt, Brace and World, 1961), pp. 47 - 78。

不包含这些术语的更一般的法则式陈述中派生出来。法则式陈述只包含纯粹定性的谓说(predicate)：陈述其意义并不要求参考任何具体对象或者时空位置。法则式陈述具有一种无限的普适性；它们并不仅仅是一种通过考察所有情形而确立的有限合取(conjunction)。法则式陈述不但是由落入其下的那些例子所支持的，而且也是由一种间接证据的联结(linkage)所支持的。

规范原则也许正是由于具有上述这些特征，才有能力核准从既往已被接受的判断派生出新的判断。伦理学者常常声称伦理原则必须只用一般性术语加以表述——而不具有那些具体的人、群体或民族的名称。这一特点也许就使得一项原则有能力核准对新情形的推理，由此使得既往的判断能够支持新的规范判断。而缺乏这一非具体性特征的概括在最好情况下也只是一种偶然的概括，而没有能力把一些数据具有的支持转移给另一些数据。当道德原则并非是原则的一个特定的*道德*方面，而是一般性的且不包含任何非定性的谓说或者具体的名称时，此一特征将有可能把数据或判断联结在一起从而支持假设推理。由此，考察道德原则的这种"形式"中有多少对于这种联结来说是*必要的*，这是很有价值的。

这里意指的是，正如我们无法把这些特征添加于偶然概括以得出科学法则那样，我们同样不能把这些特征添加于较弱的概括以得出履行推理功能的道德原则。人们可以认为，科学规律与道德原则之为真，是脱离于我们加上的任何构建的，也是脱离于我们对它们的任何使用的，亦即它们独立的真值才使得这种使用是可能的。尽管如此，诸如一般性、非专名、无地点谓说等这类特征并不是专属于*道德的*，而是对法则式特征，即对于任何东西成为(无论是科学的还是道德的)法则所必要的。在适当的语境之下，并非专属于道德的特征也能够具有道德影响。 6

人们去寻找原则,可能不仅仅是为了检验自己的个人判断,或者给予它以更多的支持,而且同时也是想去说服其他人或让其他人更加确信该判断。为了做到这一点,他就不能够只是简单地宣告他对某个立场的偏爱,而是必须要提出能说服他人的理由。理由有可能是非常具体的,但它们也很可能是很一般性的考虑,这些考虑能很好地适用于广泛情况,并在此场景下能得出一个具体判断。如果这些判断在其他情形下是其他人已经接受的,那么这个一般推理将会把这些情形征引过来作为证据,从而支持对此一情形所提出的判断。由此,原则或者一般理论就具有了一种人际间的智识功能:向另一个人证成(justification)。依据一般原则的证成会在两方面具有说服力:一是原则的表面吸引力;二是征引其他已接受的情形来支持此情形中所提出的立场。[①]

我用法官来阐释原则所具有的这种检验功能与支持功能。我设想法官的目标是要获得某一具体案件的正确判决,并且她还将过去的判决本身视作(大体上)正确的。也就是说,我把这个法官看作与道德推理者具有类似的结构。后者也要在新的场合或情境下确定什么是正当的或许可的,并且她对其他实际或虚拟的情境之下什么是正当的或许可的知识,也会被运用来构建、检验且支撑一个道德原则,据此原则可得出有关新情境的结论。

当然,法官同时也是制度结构中的人物,她做出的符合先例的、有原则的判决可能在该制度之中具有特殊的意义。法律学

---

① 通过征引其他已被接受的例子作为支持,抽象的原则性推理能支持一个具体的立场。有些论者表示,这一抽象的、非个人性的模式只是一种特殊的证成模式而已。

者告诉我们，尊重先例，也即遵循先例的原则（*stare decisis*）——使人们能够更准确地预测法律系统中的未来判决，由此他们能对其法律后果具有一定把握，且基于此来安排他们的行动。①要达到这种效果，先例不必一定是一直正确地判决的，或是为了要正确地判决这一目标而被遵循的；它们是为了产生意料之中的判决而被遵循的。其次，有原则的判决之所以可取，可能是为了约束法官的判决基础。这里要排除的是法官个人的偏爱或者偏见、当时的各种情绪、对两造之一方的偏袒，甚或是她深思熟虑的个人性道德或政治原则。人们认为，法官自己的看法、偏好、甚或深思熟虑的观点也应该如同其他人的看法一样不起任何作用——也就是说法官并没有得到一个制度性地位来使自己的个人偏好产生影响。判决在原则上应符合先例，这一要求也许是约束这类个人因素的一种手段，以限制这些因素的作用或者将它们完全排除在外。 7

然而，与科学（该领域中的目标是真理和正确性）的类比使人们对上面那个强主张疑虑丛生。符合科学数据是一项要求，但它却并不能够唯一地决定某一个法则式陈述[暂且不论界定“最优契合”（best fit）的各种方式之间所存在的出入]。总是有无限的曲线能够契合于任何的有限数据点集；其中还不止一个法则式陈述。因此，有必要找到一些其他的准则来挑选：哪个法则式陈述是我们暂且接受并要在预测中加以运用的。这些准则包括：简明性、与相关领域中已获支持的法则式陈述的类比性、②

---

① 我并没有去查证，对于人们的决策，存在着何种经验研究能支撑由法律理论者所提出的这个经验性主张，有何种其他的法律结构起到这种控制作用等。

② 我们具有的这几个数据点看来全部都在一条直线上，除了我们已经发现那样一种线性关系并不成立的相关现象。我们恰好具有这种具体数据，也许这是纯属偶然的。

与其他已被承认的理论的契合性、解释力、理论上的成效，也许还包括计算上的容易。① 单单要求一项预测按照某种法则式陈述而与既有数据相符合，这还无法唯一地确定该预测。那么，仅仅要求一位法官对一个新案件的判决要按照某一原则而契合于那些先例，由此就足以唯一地决定该判决，这种可能性又有多大呢？确实，我们发现法官们热衷于运用各种额外的准则来判案，其中也包括各种“形式”准则。② 同样，我们也可以向伦理学提出类似的问题。奎因认为，(有可能的)经验数据的全体性(totality)都无法唯一地决定一种解释性理论。各种正确的伦理原则是由对具体情形(实际的与假设的)的所有正确判断所唯一确定的，还是其间充斥着欠确定性(underdetermination)呢？道德原则除了要契合于诸具体判断外，是否还必须要进一步满足某些其他准则呢？

把原则是用作获得正确决策的工具，还是用作约束不可欲的或无关的因素(比如说个人的偏好)的影响，这两者之间是有关联的。我们做决策或判断时，想考虑的是与具体情形相关的

---

① 参见 Thomas Kuhn, “Objectivity, Value Judgment and Theory Choice”, (载于 *The Essential Tension* [Chicago: Univ. of Chicago Press, 1977], pp. 320 - 339)一文中的诸要素清单；另请见 W. V. Quine and Joseph Ullian: *The Web of Belief*, 2d ed. (New York: Random House, 1978), pp. 64 - 82。这种额外准则的必要性可能不仅是源于我们所掌握数据的有限性。奎因就声称，所有可能的观察数据的整体也不能唯一地选定一个解释理论。(参见他的“On the Reasons for Indeterminacy of Translation,” *Journal of Philosophy* 67[1970]: 178 - 183，以及“On Empirically Equivalent Systems of the World,” *Erkenntnis* 9 [1975]: 313 - 328。)没有一种恰当的解释理论，没有一种解释关系可能包含何种详尽结构的恰当理论的话，就很难确定奎因的这个强主张是否为真。

② P. Atiyah and R. Summers, *Form and Substance in Anglo-American Law: A Comparative Study in Legal Reasoning, Legal Theory, and Legal Institutions* (Oxford: Oxford Univ. Press, 1987).

所有理由且只有相关的理由。一般原则迫使我们检查其他实际的或假设的情形,以此帮助我们检验：我们认为在当下情形下相关的或结论性的(conclusive)理由 R 是否确实如此。理由 R 在其他情形下也是相关的或结论性的吗？理由若是一般性的,那么我们就能通过考察其他的情形来检验理由 R 在当下情形下所具有的力度。不仅如此,使用一般原则来决策可以使我们注意到其他的相关理由,而那些理由是我们在当下情形中未曾注意到的。检查特征 R 在其中并**不是**很有力度的其他情形,或许会使我们注意到当下情形所具有的另一个特征 F,并且正是 R 与 F 一起才具有那么大的力度。(如果我们并不曾去考察其他的情形,那么我们或许认为单独 R 就足够了。)把所有相关理由包含进来,也许还会有助于确保只使用了相关的理由,因为**如果**这些理由填满了空间,那么就会挤掉那些并不相关的理由。难道我们真的愿意接受一个不相关的理由对此一情形所施加的那种也作用于其他的情形与例子的影响吗？请注意,若我们使用假设情形或其他实际的情形来检验对当下情形的判断,这本身就已经假定了理由是**一般性的**。若我们假定事情的发生或成立是因为某一理由(或原因),并且这个理由(或原因)是一般性的,那么我们就能构建出一个能把捉(capture)这一理由的(也许是可挫败的)一般性原则,且用它来解释,科学家所研究的事件为什么会发生,为什么关于某一情形的具体判断是正确的。[①]

8

原则可帮助检验我们的判断,控制那种把我们引入歧途的个人性因素,从而指引我们在具体情形中获得正确的决策或判

---

① 理由(reasons)与原因(causes)二者的结构性特征之间的这种并列能够扩展到多远,并且为什么这一并列是成立的,这样的探查是很有意思的。“理由”也表明与概率式因果现象相并列吗？

断。基于这种观点,原则保护我们避免犯的错误乃是个别性的(individualistic)(这个情形中的错误判断)或加总性的(aggregative)(这些情形中的错误判断,一个个地是错的)。然而,这种判断放在一起的话,还可能使我们犯下另一种错误,即一种**比较性**(comparative)错误,也就是当各种情形本该以相同方式判断但却以不同方式去判断时所发生的那种错误。"类似情形类似处理"一直被人们奉为一项(形式)正义准则,这个一般性准则并未规定何种相似性是相关的。[①] 原则的作用也许是要避免这种不正义或不一致性(disparity),即不仅仅是让每一个情形本身得到正确的判断,而且使得相关的类似情形得到类似的判断。不过,假如我一直每隔两周看一次电影,我并没有必要基于类似的基础决定看哪场电影,那么,这两个类似的决策显然**不会**看作必须类似地判断的类似情形(前一个决定有可能影响后一个决定,但并不会制约它)。那么是什么堪定了形式正义准则可运作的领域呢?电影爱好者(在两个场合下)决定要去看哪场电影,他并不会将此视为在该场合下要获得**正义的**决策。比较性不正义的问题只有在涉及个别性的正义或不正义的语境下才会出现,无论我们如何标识该语境。如果情形 A 要求一种正义的决策,它得到了错误的处理,那么这一结果是坏的;现在如果情形 B 与之是相关类似的,得到了不同的处理(也就是说,正确的处理)。**如果** B 情形中的处理为这个世界引入了另外一种坏——此"坏"并非指情形 B 中的结果本身,而是这两种情形得到了不同处理这一种比较性的"坏"——且这一种"坏"超越了 A

---

① 参见 Herbert L. A. Hart, *The Concept of Law* (Oxford: Clarendon Press, 1961), pp. 155 - 159;以及 Chaim Perlman, *The Idea of Justice and the Problem of Argument* (London: Routledge and Kegan Paul, 1963)。

得到错误处理这一结果所涉及的那种“坏性质”的话，那么援引形式正义的准则，这种正义的语境就是一种比较性语境。[①] 于是，各种原则的功能之一可能就是防止这种特定类型的不正义，确保类似的情况将得到类似的判断[是同样错误地判断两种情形(避免比较性不正义)更好，还是把其中一个情形判断正确(避免了该个体情形下的不正义，但是招致了比较性不正义)更好，这很可能要取决于不同情境与情形的具体特征]。 9

人际间功能

若面临偏离原则的各种诱惑或引诱的是个有原则的人，我们仍然能够指望他坚持自己的原则。尽管他不是在面对任何可能的诱惑或者极其巨大的引诱时必然都会这样——即使这样，原则还是能阻止个人追随一时的欲望或兴趣。个人的行为原则由此就具有了一种人际间功能，亦即使得他人确信你(通常)能够抵制诱惑；原则还具有一种内省功能，即能够帮助个人自己去克服诱惑。

我们先来考虑人际间功能。若某一行为是某个人的原则指

① 我已经说过，援用形式正义准则的一个必要条件是要有这样的语境，其中要达成的决策是正义的决策，但是我并未声称这是一个充分条件。如果存在涉及正义的个体决策不具有那个比较性方面，那么就需要找到一个进一步的准则来指明，到底是涉及正义的何种语境才可以援引形式准则。在 *Anarchy, State and Utopia* (New York: Basic Books, 1974)一书的第 7 章中，我提出了一种分配正义理论，即资格理论，它明显不是一种模式理论，并且它不涉及在不同人的持有物之间的比较。然而，这并不是说形式正义准则就不能适用于人们按照相同的一般(获取、转移与纠正)正义原则而产生的持有物。因此，只要资格理论成立，除了某些人的某些持有物不是通过这些正义原则的运作而得到是一种不正义之外，如果还有某个人的持有确实是如此产生的话，那么这里还存在一种比较性不正义(例如，后面的这个人就会遭受那些不让这些持有之正义原则适用于他的人的歧视)。

令他去做(或不去做)的,我们便更能指望它。有原则者的行为是相当可靠的,故我们可以去做这样的行动,它们的成败取决于有原则者的特定行为。即便将来会有偏离的诱惑,我们还是可以信任他不会这样做,由此我们可以信赖他来安排与实施自己的行动。要不然我们就会做出不同的行为,因为该行为一无所得或变坏的风险太大了。对于与我们过从甚密的人,我们可以信赖他们的感情和一贯的良好动机来产生协作行为;但对于与我们较为生疏的人,我们则指望他们有原则的行为。

在契约法的讨论中,这类考虑是司空见惯的。契约使个人自己受到约束来实施某一行动,以此鼓励他人依赖此点并因而也完成救前者出困境的行动,但如果前者未能如此行事,则这个行动就会失败。既然第一个人要从第二个人的行为当中获益,而如果第一个人没有依约而行动的话,那么第二个人也不会做相应的行动,这样第一个人就愿意在该情形中事先限制自己的行动,即使他未来的激励有所变化。原因在于,若他的行为取决于未来变幻莫测的情势,那么第二个人就不会履行第一个人现在希望第二个人去做的那些配合性行为。

10

原则构成了一种形式的约束:我们约束自己来按照原则的指令行事。其他人可以信赖这种行为,并且我们也能从他人的这种信赖之中获益,因为他们由此愿意承担的行动能够增进我们的社会便利与协作,而这也将有利于我们实施自己的个人计划。[①] 宣告原则,把各种条件弄得一清二楚,使得人们能够更容易察觉对原则的背离,这是一种能产生(经济学家所谓的)名誉

---

① 其他人能够指望我们遵守某种原则,这也许可以阻止他们做出某些行动,而不是导致他们与我们合作。将睚眦必报作为其原则而不顾他们直接利益的民族或个人,可能会阻止别人冒犯他们。为了确保不出现例外情况,宣称这样一项原则就提高了例外情况的代价。

效应(reputation effect)的方式。对于那些要与很多人进行反复交易的人而言,这种效应是相当重要的;其他人将确信他会按规矩办事(为了避免在合作中对他有用的声誉的贬低)。①

这些考虑只能使个人想要自己在其他人**眼里**是个有特定原则的人,但是为什么他实际上想要这些原则呢?对绝大部分人而言,要使我们看上去是个有原则的人,拥有原则也许是最可信也是最简单的途径,尽管小说里和现实生活中充斥着高超的骗子。假定个人确实想具有一个特定原则,而不仅仅是看上去这样,因为实际具有这个原则将是对其他人最有说服力的方式,并且对他自己也是最容易的方式。他**能够**仅仅因为一项原则具有这种有用的人际间功能而具有该原则吗?他不是必须相信这一原则是**正确的**吗(因此,智识功能不是在人际间功能中也起作用了吗)?

若某人告诉我,他认为他持有一个原则是因为这对于让我和他人安心是必要的,则我又如何能够安心呢?我想知道的是:"你**确**实持有它吗?""有多坚决呢?"如果他仅仅将原则视作是为了让他人安心,即使这是一种非常必要而且极其有益的安慰,我难道不会想他在面对即时的引诱或诱惑时不能一以贯之吗?我认为我想要的是,那个人要相信该原则是**正确的**和**正当的**(right)。当然,他现在这样认为仍是不够的,他的信念必须是稳定的,不会因芝麻大的反驳意见或诱惑就被推翻了。这才是能让我充分地安心的东西,这样我才甘愿冒险去做成败取决于

① 美国政府想发行债券并且承诺不会通货膨胀,但一旦他人购买债券后,政府进行通货膨胀就是有利的——他人事先已经意识到了这一点。因此政府试图承诺某种货币管理规则,且让独立于国会的机构来执行这种规则,而不是把绝对自由裁量权留给自己。参见 Finn Kydland 和 Edward Prescott, "Rules Rather than Discretion", *Journal of Political Economy* 85 (1977): 473 - 491。

他的良好行为的行动。同时我也许非常善于侦察，看他对该原则是否具有真正的信念，一旦发现他不具有，我就不愿去冒合作的风险。①

11 那么，相信个人所遵循的原则是正确的，也许是她所拥有的一种有益的品质，因为这种品质使得与他人进行大范围的合作和互动成为可能。即便"正确的原则"这一概念本身是毫无意义的，这个信念依然是有用的。因为这一（让我们暂时假定）无意义的信念（由个人自己证实且由其他人探察到）将是对她未来行动的可靠指示器，它也会促使其他人做出让人信任的行为，从而有益于她。（同样，相信某个行为是神圣地规定的，所有背离的举动都将受到严酷惩罚，那么不管这个信念的真假如何或者是否有意义，只要它能向他人保证个人行为的一贯性，人们具有这种信念就是有用的。）这就提出了这种可能性，即不是针对具体的行为模式，而是针对一种客观的道德秩序的信念来做出社会生物学解释。**正确性**（correctness）的信念也许是被选择的。（对**义务论**原则的信念也许是服务于类似的人际间功能而由此被选择的吗？）

要他人确信我未来的作为，仅仅宣告我有原则是不够的；其他人可能需要不时地看到我实际上坚持了这些原则。然而，我认为是最正确或者最恰当的那些原则，它们在实施中可能很难让人观察到。最恰当的原则也许要回应很微妙的语境细节，即那些不为人知且无法被可靠地查证的历史、动机或关系上的细微差别。可以说，这里要求的不仅仅是行正义，而且行正义要为

---

① 除了把原则接受为客观有效的之外，列举并且比较一下我们还能有什么样的基础去信赖一个人的行为，这也会是有用的。这些就可能包括原则的（被列出的）其他功能，而这些功能的成功实现并不取决于要相信原则是客观有效的。

人所看见。然而,当(完全)恰当的正义所要求的回应过于复杂而无法被可靠地观察和认识到的时候,应该怎么办呢?要使他人确信正义正在被实施或者原则正在被遵循的话,那么这种人际间的功能也许使得有必要遵循这样的原则,它们不是那么精致和细微,但其应用(和误用)有时候能够为他人所查证。①

由此,原则既要精细地回应情境,也要产生公共信任,这两者之间就可能会存在着冲突。原则越精细,其应用就越难以为他人所查证。从另一方面看,当一项原则超出了一个粗糙点(a point of coarsening)之后,那么它将不再能够唤起信任,这并不是因为此项原则不能被查证,而是因为它的应用不再被认为是可欲的了。曾有人声称(这个问题是存在一定争议的),女人在道德判断上比男人能更精细地回应情境细节、关系与动机上的细微差别。② 男人和女人的这一区别(如果确实有的话)也许可被这样一个统计事实所解释,那就是女人很少在家庭领域之外做出(或想做出)决策,而在这些领域内决策的基础或动机都是可疑的。如果个人在某个(公共的)领域内必须给其他人以确信,那么在该领域内任何人都需要(多少)臣服于什么东西能够提供这种确信的指令,而原则就是这样一种工具。一旦大量女性进入到过去一直属于男性的竞技场(出于许多理由这是一件

12

① David Kreps, *A Course in Microeconomic Theory* (Princeton Univ. Press, 1990), p. 763. 他的报道是,Robert Wilson 主张,为了使潜在的投资者确信审计者自身不会被它所审计的公司所收买,进行外部审计的那种上市会计事务所就必须遵守那些已有的审计规则,而这些规则的适用是能够被外在地查证的,即使这样做不能提供有关公司财政的最全面信息。正是因为这些既有规则的应用情况是能够被查证的,所以会计事务所才能维护它作为独立第三方的那种声誉。

② 参见 Carol Gilligan, *In a Different Voice* (Cambridge, Mass.: Harvard Univ. Press, 1982);也参见 Bill Puka, "The Liberation of Caring: A Different Voice for Gilligan's 'Different Voice'", *Hypatia* 5 (1990): 58-82。

好事),道德会受到多大的影响,产生多大的变动,对此有着许多预测。但我们并没有办法确定,是这一竞技场本身的变化大,还是进入其中的女性变化大。

其他人的原则使我能够相当准确地(尽管不是完美地)预测到他们行为的一些方面,并因而引导我去信赖这些方面。不过对于此人而言,其原则对于他自己主要不是一种预测工具。人们极少试图**预测**自己的未来行为;他们通常只是去**决定**做什么。相反,个人的原则是在产生该行为中起作用的;他根据原则**指导**自己的行为。我对他人原则的所知影响到的是我对他以某种方式行为的可能性的估计,亦即我对他将以该种方式行为之概率的估计。**对于他本人来讲**,原则影响到的不(仅仅)是对于某种概率的估计,而恰恰就是这个概率本身:原则不是他将如何行动的证据,而是帮助他确定他将要(决定)做什么的工具。[1]

个人性功能

除了社会互动问题外,行为原则具有一种个人性(或一种智识)功能,正是因此,它们才能够履行和实现它们的人际间功能。[其他人(错误地)认为原则确实为某些人履行了某些个人性功能,这也许就够了。]若没有这种个人矩阵的基础(使得他人确信我们在面对诱惑时的行为,并由此引导他们选择去做与我们合作的行为),这种人际间功能是不会(作为合作博弈里的解法)出现也得不到维持的。那么,什么是原则的个人性功能和内省功

[1] 依循哲学传统,我——正如在"决定论"(determinism)中那样——使用决定(determine)这一术语意指的是:确定(fix),引起(cause),使发生(make happen),但是也请注意该术语所具有的估计(estimate)/证据性的(evidential)/认知的(epistemological)方面,正如在"我还没有确定(determine)他打算做什么"中那样。

能呢？它们又是以何种方式获得这些功能的呢？

原则也许是个人用以界定自己**身份**的一种方式："我是一个具有**这些**原则的人。"更进一步讲，长时期遵循某些原则乃是个人将其不同时期的生活整合起来并且使之更融贯的一种方法。有些人可能说有原则是件好事，因为这是使人前后一致的方法。然而，如果行动本身或行动之间是(在逻辑上)不一致的，比如某天去看电影与那天不去看电影，那么要做所有这些行动就是不可能的，并且也就不需要原则来避免这种不一致性。在逻辑上有可能一起做的那些行动之间，遵循原则也不会进一步增加逻辑一致性。行动**能够**与原则不一致，并派生地与符合该原则的那些行动不一致。但若个人只是想要避免这种不一致性，那么他根本**不要任何**原则就可以做得到这点。尽管如此，原则确实能把个人的各个行动编织在一起。通过这些原则，个人的行动和生活就会更为融贯，成为更为有机的一个统一体。而这一点本身可能就是有价值的。

13

依据原则来界定自己或身份指的是什么呢？我们应把自我设想为一个原则体系吗？这些原则可能包括了改变既有原则和整合新原则的原则，因此这也是依据原则改变自我的原则。(如果个人违背了自己的原则，那么这不是有毁其自我之威胁吗？)但是连贯的目标也能整合个人不同时期的生活与行为，那么为什么要用原则而不是目标来定义自身呢？个人不用原则来**定义**她自身，但仍然是可以**具有**原则的，只是不是作为其身份的一个内在的组成部分，而是作为对以一个独立可区分的身份之行动的一种外在约束。人们会想到康德的自我创造(self-creation)与自我立法(self-legislation)的论点。然而如果选定的目标能够自我创造，那么为什么还需要自我**立法**呢？原则的这一作用要依赖于康德的"什么(并且只有什么)

引起自主性自由(autonomous freedom)”这个有争议的主张吗?

原则的这些个人性功能关注个人作为整体的(或至少它的很大一部分的)生活或者身份。原则也在微观的层次上更为温和地对一个人起作用。道德原则的内省功能就是与我们对这些原则的承诺有关联。当我们开始做长期计划时,就会产生一个问题,即将来我们是否还会坚持它们,或者如某些人会说的,我们的未来自我(future self)是否还会执行它们。只有当答案是肯定的时候,开启某一特定计划才是有价值的,且只有当我们对它得以维续有一定把握的时候,开始实施它才是合理的。如果现在我把某些东西*作为一个原则*,就会使得将来背离此原则的代价更大——如果不是背离原则的话,则相同行动的代价会更小——那么如果一个计划纳入了一个现有且会长期存在的原则,则我放弃此计划的可能性就更小。这并不是因为我还有另外的原则促使我坚持自己的计划,而是因为这个计划本身体现了我(很可能)会继续持有的一项原则。正如同原则具有给予他人以信心(她在做自己的计划时能指望我的行为)的人际间功能那样,原则还具有这种内省功能,使我自己能够指望未来的自我——当他很可能也将持有该原则时——的某个行为。因此,我现在才可以合理地采取这样的一些计划,它们的可欲性要取决于我未来的某个行为。

14

在个人决策的过程中,原则也可能发挥一种排除性或过滤性工具的作用:在选择情境下,不把那些违背原则的行动列为备选项。因此,原则也会为一个“有限理性”的生物节省决策的精力和计算的时间。然而,这种排除也不一定是绝对的:如果在备选项中没有足够好的行为(指高于某一期望水平),那么就可能会重新考虑先前被排除掉的选项。

克服诱惑

我想集中讨论原则的主要内省功能,亦即使我们克服诱惑、障碍、注意力不集中和分心等。心理学家安斯利(George Ainslie)提出了这样的理论,解释为什么我们会去做明知违背长期利益的那些冲动行为,我们可用什么手段来对付这种诱惑。① 在我们转向安斯利的工作之前,先了解一些背景情况是有益的。

经济学和心理学数据显示,相对于在日后报偿最终得以实现的时刻而言,我们在当下对于一项未来报偿总是不那么在意:我们会对未来进行"打折"。未来收到的一项报偿对于我们的现时效用(current utility)小于这一报偿在日后实现之时的效用;收到回报日期越遥远,其现时效用就越小。这种情况本身是一种有趣的现象,我们可以对它的合理性进行质疑。在我们的行动计划及项目规划中,任何时候我们不都应当如同报偿实现时那样来评估它吗?诚然,我们也要考虑不确定性,即我们是否会活到报偿实现之时和报酬是否能够实现——每个事件可能都不是完全确定的。那么,在我们目前的计算中,我们想用的期望值是用概率对未来报偿打过折的。但是,实际得到该项报偿的效用不是应当无论何时都保持不变吗?

时间偏好(time preference)——一些经济学家用来表示未来效用折扣的术语——也许是一种进化方式,由此对那些不能

---

① George Ainslie, "Specious Reward: A Behavioral Theory of Impulsiveness and Impulse Control", *Psychological Bulletin* 82 (1975): 463 - 496; Ainslie, "Beyond Microeconomics", in *The Multiple Self*, ed. Jon Elster (Cambridge: Cambridge Univ. Press, 1986), pp. 133 - 175.

15 做出这种预期概率计算的生物灌入一种有大致效果的机制。天生的时间偏好可能是一种粗略的经验规则，它近似于先前的计算会产生的那种行为或决策，至少就报偿（及惩罚）会影响到全面适合度(inclusive fitness)而言；可能一直就存在着对这种时间偏好的选择。① 那么，若一种存在物具有认知工具来明确地考虑未来报偿的不确定性，还能对未来实行明确的概率折扣，则这里就出现了一个问题。如果我们已经被植入了一种天生的时间偏好——替我们的祖先计算这种概率折扣的进化意图——并且不仅如此，如果在我们的概率计算中所明确打折的乃是未来报偿（已经通过时间偏好打折过了）的当下价值，那么，这里所发生的就是一个**双重打折**。无疑这折得太多了。看来对那些足够老到而认识到这一切并且做出期望值计算的存在物而言，他们应该使用未来报偿在其实现之时的效用作为当下的估计（然后，这个效用会经由概率而被明显地打折），而不是使用未来报偿因受时间偏好影响而被折扣过后的当下价值。要不然，他们就应该跳过预期价值的估算，直接坚持以进化方式植入的时间偏好。② 然而，若纯粹的时间偏好本身即是一种合理的现象，而不**仅仅**是对概率折扣的一种进化替代物，如果这种进化的塑造作用确实发生了，那么情况就更为复杂了。

描述对未来报偿的时间偏好折扣的曲线，不一定要用直线

---

① 我们能使用有关人们当下的时间偏好程度的信息，从而对这个“偏好度”最初于其中进化的那个生物的环境险恶和生活史做出一种粗略估计吗？我们能用关于时间偏好曲线之总体形状的信息，从而来检验选择在其中起作用的领域的理论吗？［比方说，在近亲选择(kin selection)中亲族有多大呢？］

为了在部分程度上接近，除了概率折算外，时间偏好可能还选择其他的特征。Susan Hurley（在谈话中）提到，由于未来的偏好改变有可能会导致效用的改变。

② 我首先是在“On Austrian Methodology”, *Synthese* 36 (1977): 353 - 392 一文中讨论了双重折扣的危险。

或者指数曲线；它们可以是双曲线的形状。[①] 安斯利注意到这样的两条呈高度弓形的曲线(就像双曲线一样)能够相交，并且他探索了这一事实的含义。(在图1中，y 轴衡量报偿的效用；此一报偿在给定的时刻对于某人的效用，则由那时所对应的曲线的高度来衡量。曲线向左侧下倾，这是因为一项未来的报偿在越早的时间点上其价值就越小。)假设有两套行动项目或计划，二者会导致不同的报偿。如果你取得其中可较早得到的那份报偿，也就是接受二者之中报偿较小的那一份，那么，这将排斥或者阻止你获得可更晚得到但更大的那份报偿。随着时间的推移，个人总是要实施所处时间点上效用最高的那个计划。在时段 A 中，更远的报偿具有更大的效用；然而在时段 B 中，更近的报偿具有更大的效用。既然实际上更多的报偿只有到时段 C 的终点处才能得到，因此若要得到那份更多的报偿，这个人就必须度过中间的时段 B 而不在 B 时段内转而去追求较小的报偿。而这就引出了一个问题：在中间时段 B 内，**立刻**就会收到较少报偿的前景比随后会收到更大报偿的前景有更大的效用。

16

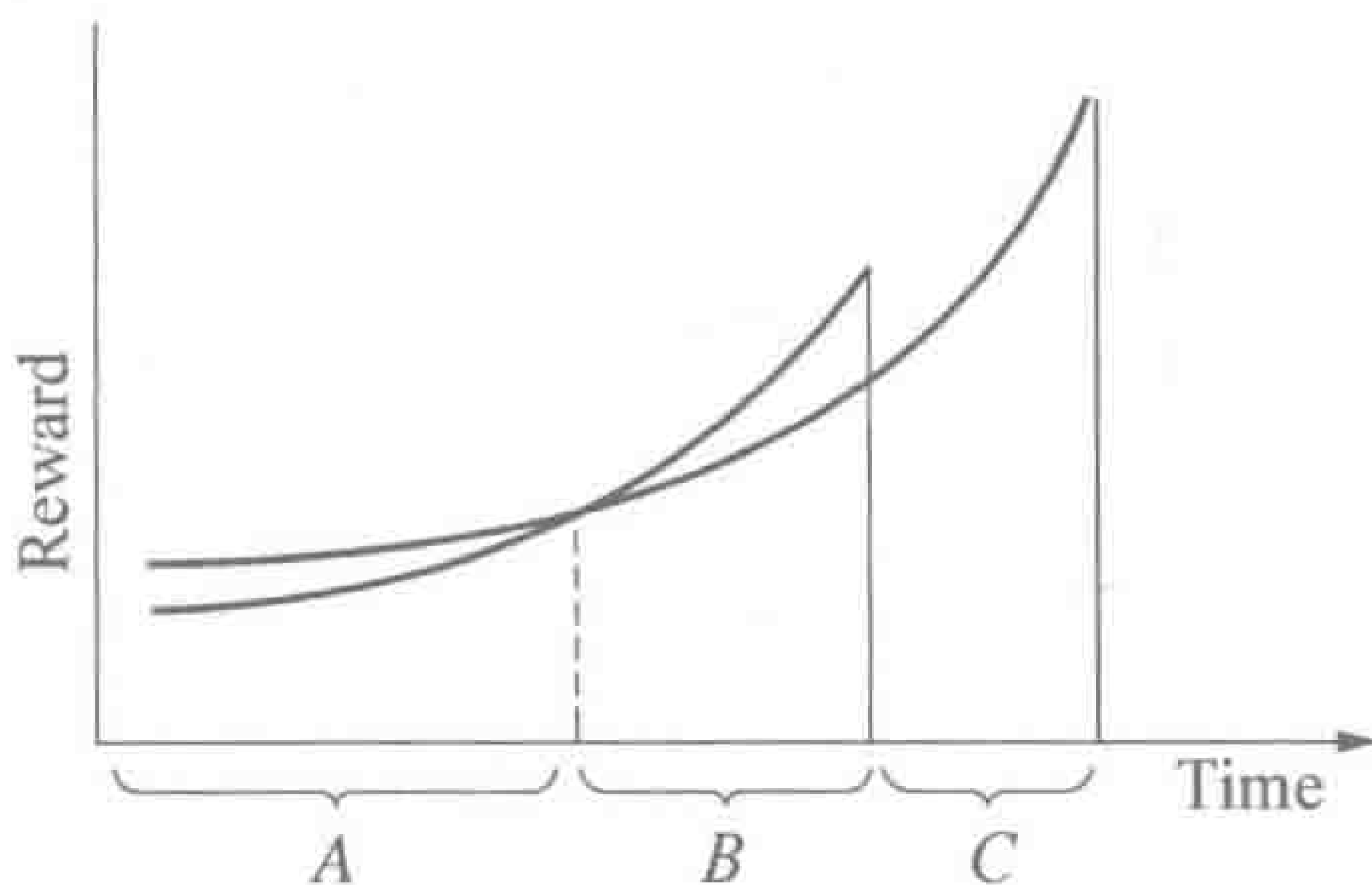


图 1

---

① 这最后的曲线形状是"匹配律"(matching law)方程的一种结果。参见 Richard Herrnstein, "Relative and Absolute Strengths of Response as a Function of Frequency of Reinforcement", *Journal of The Experimental Analysis of Behavior* 4 (1961): 267 - 272。

我们为什么认为这个人*应该*尽力克服那一中间时段呢？为什么她不应该选择更小但是更快的报偿呢？[①] 是什么使得 A、C 时段（那里较大的报偿突然变为有最大的效用了）成为决定何种选择为最优的适当时段呢？在这两个时段中，行为人会倾向于取得那份较大的报偿，而在 B 时段内，她却倾向于获得那份较小的报偿——也就是说，她此时获得的报偿小于在另一种选择中她会得到的报偿。而当我们说避免这种诱惑是更好的选择时，*我们*所处何处？为什么那个立场要比时段 B 内的立场更恰当呢？

这里有一个建议。时段 B 之所以不是决定那个人应当做什么的合适的基准点，是因为时段 B 并不是她对于这个问题看法的代表性样本。（因为）时段 A 和 C 相加的总时段比 B 更长。另外，若我们补上她在报偿被实现的时刻*之后*的判断，并用图标明*那时*对她而言何种报偿最大，我们就会发现，她在消费掉那一较小的报偿之后，很快就宁愿自己不曾这样做过；但是在消费掉那一份较大的报偿（在时间段 C 的终点处）之后，她会继续偏好所选的较大报偿。我的意见是，克服诱惑和得到更大报偿通常是更可取的选项，原因在于这是个人在大多数时间内的偏好：这是她（相当）稳定的偏好；而另一个选项只是她在一个非代表性时刻内的偏好。[②]（不考虑任何事后偏好，如果时段 B 持续的时间比时段 A 和 C 要长，*在那种情形下*应该抵制该诱惑还是显

---

① 我要感谢 Amartya Sen 提出了这个问题。

② 这里面还有一个*后悔*的现象，即由于现在回顾不可欲的过去行为而致使当下效用的降低。具有一种后悔倾向可能多少有助于你克服 B 时段的诱惑，因为在 B 时段你能预期到，如果现在选取了那一较小且更切近的报偿，那么在 C 时段及其后的时间内这一报偿的效用水平会降低。但是这种预期能够充分地反馈进入到 B 时段内的总体效用而影响当时的决策吗？

而易见的吗?)诱惑并不总是应该被抵制的,只有当对于更大报偿的向往(包括事后的偏好)乃是个人在大多数时间内的偏好时,情况才是如此。这一准则肯定是可以被挫败的,而不是结论性的。但是,相对于下面两种说法而言,上面的那个准则确实具有更贴近人的偏好这一优点(尽管它并不是嵌入在某个特定的局域偏好中),即这仅仅是人出于利益的考量而去追求后期更大的报偿(因为这种实际的报偿),并因此去抵制这种诱惑,或者相关的准则——并且抵制诱惑服务于——最大化个人一生的效用。①

17

安斯利描述了各种各样能让人摆脱那个充满诱惑的中间时段的方法,其中包括:在时段 A 中采取一项行动,使自己无法在时段 B 中追求那一较小的报偿[例如,奥德赛(Odysseus)把自己绑在桅杆上];在时段 A 中采取一项行动(也许是和另一个人打个赌),这会在你获取那一较小报偿时给你加上一种惩罚,由此改变它在时段 B 中的效用;在时段 A 中采取一些措施以避免自己在时段 B 关注或者念念不忘那一较小报偿的诸般好处;②还有就是——我们本章的主题——构建出个人的一般行为原则。

行为的一般性原则会给行为分类;它能把某一具体行为与其他行为归为一类。例如:“不要在正餐之间加餐”;“不要再吸一根烟”。[有人可能认为原则比规则(rules)更为深刻,也没有那么机械——这是在法哲学文献中常见的区分——但是就当前的目的而言,我对二者不作任何区分。]我们可以试着在效用理

① 对于最大化终生总效用这个单一目标的批判性讨论,参见我的 *The Examined Life* (New York: Simon and Schuster, 1989), pp. 100 - 102。

② 另请参见 Jon Elster, *Ulysses and the Sirens* (Cambridge: Cambridge Univ. Press, 1979)。

论与决策理论中把对行为的原则性分类效果描述如下。原则会把一些行为全部划归为类型 T,并类似地对待它们,由此把这些 T 类行为的效用(或说是这些行为结果的效用)连接起来。但是因为这一原则的存在,所以一切 T 类行为一定会产生相同的效用,这种说法就太强了。一个特定的 T 类行为可能同时也属于其他的类型或原则,其他的 T 类行为则不,因此,它们的效用就可能是不同的。原则所建立的是受此原则所统摄的各种行为的效用之间的那种**关联**(correlation)。这一点在偏好的层次上表述就是:若 T 类行为与其他行为在偏好序列中排序,则 T 类行为的排序序列之间具有一种关联性。然而,若这种关联性就是采纳或者接受某些原则而对行为效用产生的唯一影响,那么,原则对于我们克服诱惑这一目的来说就毫无帮助。

一项原则("不要在正餐之间加餐";"禁止再吸一支烟")的标志在于:是否把去做一件即时的具体行为的决策(吃**这顿**快餐,吸**这支**烟)与原则使得该行为所属的整类行为连接在一起了。现在,这一行为就代表了整类行为。通过采纳某些原则,你似乎就使得这点为真了:如果你做了这类行为之中的一个行为,那么你会去做整类行为。如此一来,筹码就提高了。这一次加餐行为具有的效用,就与未来所有那些加餐行为所具有的负效用或说害处捆绑在一起了,而这就有可能帮你度过那一个诱人的 B 时段;这次特定的加餐,现在对你的效用也就改变了。这次的加餐变成了能代表所有加餐的行为,并且在这一初始时刻,保持日后的苗条身材和身体健康对于当下的效用,远远大于那些未来加餐的快感之于当下的效用;臃肿的身材或者差的身体对于当下具有的负效用,成为当下考量的这一具体加餐行为

18

的一项特征。①

但我们为什么认为,这个人会在时段 A 而不是时段 B 构建出一项原则呢? 这个人为什么不在这次决定加餐,且构建出一项始终都加餐的原则呢? 或者更一般地说,为什么他不构建出一项总是屈服于即刻诱惑的原则呢? 但是构建并接受这样一项原则(与现在加餐这个行为一起)本身并不会带来直接的报偿或者使不同时期的报偿最大化。它一般只是减少报偿的延迟。但是在 B 时段,即面对一项具体的诱惑时,我对任一个报偿都**总是**想"减少延迟"吗? 不是的。因为尽管就一个特定的报偿而言,我是处于 B 时段,但是就众多其他(成对)的报偿而言,我是处在 A 时段(或 C 时段)。就其他这些更为遥远的成对的大报偿与小报偿而言,我现在并不总是想去获得那些更为直接的报偿,即使由于我处在**它所在**的 B 时段而确实希望得到一份更直接的**特定**报偿。正是因为诱惑是历时性地分散的,所以在任何一个时刻,我们都有比 B 时段要多的 A(或 C)时段。因此我们不会接受一项总是臣服于诱惑的原则。②

---

① 此处着重关注的是个人领域内某类行为的全体,这可能会使有些读者想起公共领域内的规则功利主义。但是相反,我们的问题是,对于一项一般原则的接受如何影响对一个个别行动的选择,而这个行动在没有原则的情况之下将不会产生最大的效用。一个可堪比较的问题是,一个抱有行为效用主义(act utilitarian)愿望的人,在(由于某种原因)依照一种规则效用主义的原则做出决策时,他如何能够在个别选择情境之下让它起作用呢?

② 倡导屈从于诱惑的人可能回应道:"你说的是我们不**想**总是屈服于诱惑。但你说过,一个原则就是让我们摆脱当前欲望的工具。那么可能我们需要的就是让我们摆脱不总是屈从于诱惑的这种愿望的原则。"暂且避开这种悖论,当在 t 时段内有一种相反的欲望比 t 时段内的诱惑更为强大的时候,一项原则在这样的情况下(最容易)会被采纳。(这一诱惑在时段 t 过后才达到最强。)不会存在这样一个时间段,即屈服的欲望总是不比相反的欲望弱(如果确实出现了这样一种短暂时期,由此采纳了一个原则,它马上就会因为后来的一个非暂时的欲望而被推翻掉)。

我们采纳一项原则会使一个行为代表许多其他的行为，也借此改变了这个特定行为的效用或负效用。此种效用的改变乃是运用我们的能力和力量，从而使得一个行为代表或象征其他行为所导致的。这一次侵犯了原则并不必然使得我们总是侵犯它：这一次加餐并不必然使得我们将成为连续加餐者。在我们采纳这个原则之前，这一次做了某一行为并不涉及到总会如此作为。采纳这个原则制造了(forge)那种关联，因此，对这一次 19 侵犯原则的惩罚便具有了今后总是侵犯这一原则的那种负效用。我们是如何确切地做到这一点的，对此的探究将是富有教益的。

我们有这样一种能力，这一事实有着重要的后果。我们能够如此改变效用(接受一项原则使一个行为代表其他的行为)，但我们不能太过频繁地这样做，也不能过于僵化。如果我们侵犯了一项已采纳的具体原则，我们没有理由认为下一次的情形会有所不同。若每次的情形都一样，我们在这一次这样做了，那我们在此类场合下不是总会这样做吗？除非我们能够区分开这个场合和随后的场合，有理由相信这个区分在今后有足够的分量，使得我们不会通过构建出另一种区分(而这是我们在随后的情形中也同样不会坚持的)而再次纵容自己；否则的话，此次所为将致使我们预期会一直重复这样做。(要构建出一个区分能够允许这一行为而排除以后的重复，实际是要构建出另外一个原则了；我们必须有更多的理由去相信，我们更会遵循那个新原则而不是这个原则，否则，这一重构也不会给我们未来的节制带来任何可信性。)在这种情境下，这一次的作为意味着我们将来会继续这样做。这难道还不够我们把所有那些未来重复之负效用附属于这一特定行为，从而改变这一次行为的效用吗？

我们预计：如果这次这样做了，则未来也会反复这样做。但是我们这一次的行为真的会影响将来吗；它会使得我们重复这一行为的可能性更大吗？还是这一行为影响到的只是我们对于重复行为有多大可能性的估计？我们需要考虑两种情境。如果我们过去没有接受过排斥此次行为的任何原则，那么依据心理学家所谓的“效应法则”(law of effect)[一个行为的正强化(positive reinforcement)会增加它在未来发生的概率]来看，这种行为对重复概率的影响不大。如果过去已经发生过许多类似行为的话，那么这对于重复概率的估计值会提高一点。但是如果过去已采纳了一项原则，那么无论是对于旁观者还是行为人自己而言，违反这一原则的行为都会提高对她重复这个特定行动的可能性的估计值。这一次违反也增加了她再去这样做的可能性。原则已经破坏了；行为的一个障碍物已被拆除了。不仅如此，认识到这一点还会使行为人气馁，她尽力在将来避免这一行动的可能性也就更小了。(请注意，如果一个行为影响到她自己对于未来相似行为之概率估计的话，这个行动会使她气馁，并因而影响到实际的重复概率。)因此，构建出一项原则对其排斥的行为筑起了一道额外的屏障，这是把全部行为的影响系于任何一个(过去的)行为。个人对一项原则投入得越多，即曾经为 20 遵从它所付出的努力越多，那么现在违反它的代价也就越大。(因为你付出了如此多的努力尚且不能成功坚守这项原则，那么你又有多大可能性来坚守另外一项原则呢？)不仅如此，这次遵守了原则也是一种服从于“效应法则”的行为：即正强化增加了在未来也遵守原则这一事件的可能性。

然而，违反一项原则的后果也许要更为普遍，因为它可能对你在任何一个领域内(当面对的诱惑与你这次所臣服的诱惑同样强烈时)是否能成功地遵守任何一个原则的概率或可靠性都

会产生影响。的确,你可以试图堪定并尽量把损害限制在这一个领域之内,但是,这与试图将损害限于这一个领域之内的这一次违反行为会面临同样的问题,只是上升了一个层次而已。当违背义务论原则直接威胁到了未来的任何一个原则性行为时,这些义务论原则就具有最大的分量:如果我(在这种环境下)连这个原则都侵犯了,那我怎么可能相信自己还能成功地信守任何一个(可欲的)原则呢?以一种过度的康德式热情,人们也许试图通过构建出一种“绝不侵犯任何一项原则”的(元-)原则,来增加灾难扩散(spreading disaster)的潜在效果。然而,即便“使任何一个违背行为代表所有行为”能降低任何一个既定违背行为的概率,最轻微的违反行为的实际后果也将会被危险地放大。这并不是说,侵犯原则的一个行为,因其代表着所有行为,则它的发生也就意味着消解掉了该原则,因而其后人们就可以随心所欲地去违反它而不受任何责罚。一个行为附带有全部行为的负效用,即便第一个行为已经做了,但其后的每一个行为也还是具有那种负效用。这种负效用不能通过违反原则而避免,只能经由抛弃原则而得以避免;但若抛弃原则,个人就面临着这个原则旨在避免的那种负效用。

采纳一项原则本身是一种行为,会影响到其他行为之间的概率关联,故而在选择采纳何种原则时慎重一点才是恰当的。个人必须考虑的不仅仅是坚持原则所可能带来的好处,还要考虑违反原则的概率和这种违反的未来效果。也许采纳一个(当遵守时)不那么好的但较为容易维持的原则是更好的,特别是当你没有坚守住更严格的原则,并非总有其他的原则可以作为可靠的退路时。(此外,你也想要一个足够明确的原则,从而使是否违反原则一目了然,这样一来,个人的未来自我就无法在原则是否得到遵守这个问题上轻易地蒙混过关。)无疑,参酌以上这

些考虑，我们可以构建原则的一种最优选择理论。①

原则一般谈论的都是一类行动的全体，并且它还使每一个当下的行动都代表着全部行动。为了发挥原则使个人摆脱诸如在 B 时段中的诱惑这一功能，谈及的必须是*所有*的某类行动。我们没有这样的原则，它说*绝大多数* P 应该是 Q 或"15%的 P 必须是 Q"。（或者，如果我们确有这样的原则，那它们也不是用来对付这相同的诱惑的。）尽管有时候我们需要做的不过是在部分或绝大部分时间做一个行为而已（比方说，在绝大多数晚上不吃饭后甜点；每月清偿我们绝大部分账单）。尽管如此，我们经由原则来实现这点的方法还是构建出一个谈及"所有"或"每一个"的陈述，与我们所想要的那种混合是可以共外延的。例如，*每*个月，清偿你的绝大多数账单；*每*个星期，在绝大多数晚上不吃甜点；*每一*年，参加全体教职工大会。一个教师——不是我自己——他的原则是在*每个*班都不会给出太多的 A 等成绩。由此，这每个星期、每个月或每个班就变成了代表全体。因此，我们能够解释：为什么旨在克服诱惑的那种原则关注的是一个类别之中的全体而不仅仅是一部分（一项*规范*本身可以关注 n%，这里 n 介于 0 与 100 之间，但一项*原则*则不能）。一项原则具有某些功能，而要实现这些功能，一个情形就必须代表或者象征全体。因此，观察到的原则的这种"全体"特性会支持上面的观点，即所说的原则具有那些功能和以那种方式实现这些功能。②

21

---

① 一项原则的颁布也会影响到第三方将其付诸实施的方式；一个原则的设计者要考虑到其他人可能会以怎样的方式扭曲或者滥用这些原则。与此相关的一个观点，即关于诸如马克思和弗洛伊德等社会理论者本应该采取预防粗俗化的措施，请见拙著 *The Examined Life*, p. 284。

② 对于原则容纳"全体"的另一种解释是：原则编码（codify）理由，而理由是普遍性的（尽管是可挫败的），因此原则也具有普遍性。但是，即使这种百分比是一样的，为什么理由不是"在绝大部分情况下"而是"普遍但可挫败的"呢？

原则看上去可能是实现我们目标的粗糙工具;它们的普遍涵盖性(要实现任务就要放弃*所有*的甜点和分心)对于达到我们的目标来说也许并非是必要的。"全体"涵盖(饭后甜点,星期)的东西具有松动的余地,多少缓和了这一点,减低了原则的过分苛刻性,但还是留下了一些缺陷。如果对一个行为的 n 次重复有一个清晰的门槛,迈过这点继续重复行动将使得目标受挫,但在此之前,目标还是能达到的,那么,一个理性的人不会恰好重复 n 次行为后停下来吗?(如果每次重复都会增加实现目标的难度的话,则需要一个更为复杂的陈述。)既然第 $n+1$ 次行为本身将产生打破平衡的后果,那么便不需要任何原则来排除它了。这可能(近似地)是一种个人何时决定停止吸烟(或增加体重,等等),并因此决定何时设置一项原则的理论。然而,既然有诱惑,它就是在**那时**需要设置的一个原则。

沉没成本

安斯利所提到的摆脱 B 时段诱惑的一种方法是这样的:在更早的 A 时段里自己承诺要追求在时段 B 和 C 中那个更大的报偿。这类"承诺"的一种模式是,在 A 时段中为(将来)追求那一更大的报偿而投入大量资源。若我认为,今年多看点戏剧或者多听点音乐对我有好处,我也知道自己常常会在上演当晚没有太大兴致出门,那么尽管我明知在演出当晚还有很多票,我还是可以事先买好很多演出的门票。因为不想浪费已经花出去的票钱,故相对于把决定留到演出当晚而言,我就会去看更多的演出。诚然,我可能不会用完*所有*的票,在有些夜晚慵懒还是会占据上风,但相对于不提前买票而言,我去得还是会更多。知道这一切,我就会决定事先买票,促使自己去看更多的演出。

22

经济学家们提出学说，认为所有的决策都应当只关注各种可选行动之(当前和)未来的结果。行动各个过程中的投资成本已经付出了。尽管现有资源会影响到我们所具有的各个行为的结果——我已经有了票，观看演出是没有任何未来花销的——并经由这些结果而被考虑进来，但在个人决策时，“他已经承担了去促进某一项目的成本”这一事实却不应该有任何分量。因为这些成本，经济学家称之为“沉没成本”，已是明日黄花；而对于现在来说，重要的只是未来的收益流。因此，若我现在更乐意待在家里，而不想跑出去观看一场(不用花钱的)演出，即今晚在家里坐着比外出看演出对我而言具有更大的效用，那我就应该待在家里。我已经在演出门票上花了钱是没有影响的——经济学家的“沉没成本应当被忽略”的学说就是这个样子的。[①]

从最大化金钱收益的角度来看，这可能是一条正确的规则；但由于耳熟能详的理由，它并不是决策的恰当的一般原则。我们并不是只在我们的未来收益受到影响时才重视自己过去对他人的承诺，因为违背自己的承诺会影响到其他人对于我们的信任，并因而影响到我们未来获得其他利益的能力；而且，我们也不会把过去在现有的工作及生活的规划之中所倾注的努力视为无足轻重的(除了继续这些规划比启动其他的新规划更有可能

---

① 正如当人们在虚拟选择之下的决策所显示的，人们常常并不遵守“忽略沉没成本”这一学说。关于这一点，参见 H. R. Arkes and C. Blumer, “The Psychology of Sunk Cost”, *Organizational Behavior and Human Decision Processes* 35 (1985): 124 - 140。Arkes 和 Blumer 把买票例子中偏离这一学说的人看作是不理性的。

Scott Brewer 怀疑，人们是否经常会这样预期：如果自己不用掉今晚这张票，那他就会担心自己将来会再去买另一张票，或者进行某些其他的娱乐花费，且(在他的预算体系里)他并不希望在娱乐上花更多的钱；因此，成本可能并不全部是过去的。请注意这位经济学者是把这种分解计算体系指责为不合理的。

带来收益的情况外)。这样一些计划定义了我们的自我和生活的意义。[1]

我们一直在讨论的这个问题还指明,把忽略沉没成本的学说作为决策的一般原则具有另一个缺陷。我们并不忽略沉没成本,这一事实为我们提供了摆脱在 B 时段选择较小但更直接报偿的那一诱惑的一种方法。在较早的 A 时段,当我们能够清楚地看出更大但更遥远的报偿有利可图时,我们可以为获得那一报偿而投入资源和精力。因为我们知道,当诱惑来临时,我们现在不想(将来也不想)浪费掉那些资源,这一事实将作为一个重要的理由来反对选择较小报偿,由此增加了那种选择的负效用。如果我知道我会在未来的一些夜晚受更小更直接的舒适报偿所惑(不用在雨夜出门等),然而我也知道现在及以后我都会为自己去看了所有演出而感到高兴,那么,我就可以现在提前买票,来激励自己在演出当晚不窝在家里。

人人都把在 B 时段屈从于较小的报偿看成是一个问题、一种不理性或是一种不可欲的短见。当事人自己也是这样看的——事前和事后,如果在当时不是的话——我们思考它时也是这样看的。经济学家也把另外一种行为方式,即重视沉没成本,视为不合理的和不可欲的。但我们现在看到这后一种行为,如果事先预计的话,能够用来限制和阻止第一种不可欲的行为(即屈从于更小但更近的报偿)。我们可以有意地把我们认真对待沉没成本这个倾向作为一种增加我们未来报偿的手段。如果这种趋向是不合理的,那么可以合理地用它来阻止且克服另一

---

① 参见 Bernard Williams 的论文,载于 J. J. C. Smart and Bernard Williams, *Utilitarianism: For and Against* (Cambridge: Cambridge Univ. Press, 1973); Williams, "Persons, Character and Morality", in *The Identities of Persons*, ed. Amelie Rorty (Berkeley: Univ. of California Press, 1976)。

种不合理性。如果某人给我们一粒药，它使得我们再也不会在意沉没成本，那么接受它是一个坏的建议；这会剥夺掉一个有价值的工具，我们可以用它来摆脱（未来）某时的诱惑。（尊重沉没成本的这种倾向可能是适应性的，即它可能是在进化过程中得到选择的吗？）因为重视沉没成本有时是可欲的（所以经济学家的一般性谴责乃是错误的），而有时是不可欲的，所以吃下这粒药是否可欲就取决于这两种对抗情境之间的数量比和筹码。

我之前提过，个人在遵守一项用来克服即时诱惑的原则上所付出的努力越多，则违反这项原则的代价就越高。如果你付出这么多努力尚且不能坚守这项原则，那么你也不太可能做到坚守另外一项原则。认识到这一点会使你很有理由坚守住目前的原则——这是眼前的那个救生筏——并因而在你面对这个具体诱惑时给予不去违背原则以更大的分量。我们成功地遵守了的那个行动归类（以避免即时的诱惑）就因此获得了进一步的韧劲。要注意这里涉及一个沉没成本现象。坚持这项原则和与之相关联的归类，其背后的推理涉及的说法是：如果我连付出了这么多努力的原则都坚守不了，那我怎么会有希望坚持另一个呢？只有当我是一个尊重沉没成本的人时，我才能做出这样的论证；只有一个重视沉没成本的人才会拥有一个理由去坚守他目前的这项原则来摆脱诱惑，而不是这次选择屈从、然后再构建出一个不同的原则，然而时机一到，这项原则也会被抛弃，并且其也许恰恰是基于第一种检验。正是沉没成本使得目前这个原则成为采取立场之地。［不要这样争辩，即对两种不同行为过程——亦即坚持当前策略与屈从于当前诱惑然后构建出一种新策略——的未来结果而言，这些是未来导向的考量，因此不尊重沉没成本的当事人也可以经历相同的推理思路；恰恰是因为尊

24

重沉没成本这个已知倾向使得一种行为过程将拥有(也是可以被预见到拥有)与另一种行为过程明显不同的后果。要不然凭什么认为,如果我现在不侵犯原则,那么我将继续遵守它,这种可能性要比我侵犯旧原则之后再遵守新原则的可能性更大呢?]也许我们尊重沉没成本这个已知现象在我们遵守已经采纳的原则上起了某种作用?我现在知道,如果我们能做到在一段时间内遵守这个原则,那么我们对这个原则有所投入这个事实就会在将来为我们(作为重视沉没成本的人)提供继续遵守这一原则的理由——而这一点现在也会为我们提供某种理由。①

对于这些由我们重视沉没成本而实现的功能而言,经济学家可能回答道:对一个完全理性的人来说,重视沉没成本根本就是不可欲的;只有某个*不怎么*理性的人,才会沉湎于此。然而,这个结论却并非如此显而易见,即使暂且不论早先所提到的因素:对他人所做出的承诺以及过去对自己的工作和生活规划的投入。若具有一种手段来说服别人,让他相信,即使我们面临着使得坚持自己的目的与规则会在未来不利于我们的这种威胁,我们还是会坚持原则——以这种方法抑制他们做出或实施这种威胁,这在人际间可能是有用的。② 即便你没有做其他不合理行为的倾向,并且你所试图去说服的那些人也没有这样的

---

① 我把这个建议归于 Susan Hurley。参照且并列于我们早前的问题,即如果某人持有某个原则的唯一理由是我们信赖他遵守原则时对他的好处,那么我们是否能够信赖他会遵守这个原则,她也问道:如果每个人早先没有独立的理由来投入成本,即不同于让他随后重视那些成本的理由,那么是否能预计他会重视他所投入的那些成本。

② 参见 Thomas Schelling, "The Art of Commitment", 载于他的 *Arms and Influence* (New Haven: Yale Univ. Press, 1966), pp. 35 - 91。也参见 Schelling 对于"不合理性的合理性"的讨论。

倾向,这依然是有用的。[①] 然而,以一种不理性来对抗或阻碍另一种不理性,这种论点是值得指出的。我们认为不合理的其他事物——也许是意志软弱、自我欺骗或者推理谬误——之中,有什么能够被有意地用来阻拦或限制其他不合理的或者不可欲的事情吗?(而且,这样一些显然不合理的倾向的全体可能会比那些显然是——分开来看的话——合理的倾向的全体更好地起作用吗?) 25

请允许我提及另外一种技巧,人们也许会用它来使自己度过 B 时段的诱惑(那里较小的报偿突显得如此之大)。她可以考虑另一个处境完全相同的人(其福利是她所关注的),看他会推荐什么样的行动,然后自己也采纳那一建议。拉开与情境的距离,不是仅仅从一个时间点向前看,而是非个人化地观察图表,也许是平息那种更近(但最终是更小的)报偿之诱惑的一个办法。这一过程要求你具有一种能力,当你身处其境时能非个人性地观察情境,并且认为这个相同的选择原则既适用于你,也适用于其他人,也就是说其他人在这种情境之下也会做与你一样的行为。对于超越交叉曲线中的 B 时段来说,也就是最大化这个人的总报偿来说,强烈倾向于这种不偏不倚的态度是极其有用的。而且正是这种倾向本身构成了一种伦理判断的成分:适用于自身行为的同一个原则也适用于其他人的行为。

原则还有一项功能是我尚未提及的:**划分界线**。原则标出了一个我们将不会踏出去的边界——"这是我划下界线的地方!"——而且我们会想,"如果我不把界线划在这儿,那我划在

---

① 对各种可能规划之成功机会持有一种乐观心态,这可能是一个有用的特质,尤其对年轻人来说——要不然将没人去尝试新的和大胆的事情,但它也倾向于坚持正在进行且已投入了大量资源的规划——要不然,在第一个严重的困难关头,人们就会转向另一个未试过且对之仍然有(过分)乐观态度的规划了。

哪儿呢？"在一系列情境(a gradient of situations)中可能再没有其他明显的位置好划线了，即在可接受的领域之内再没有明显的位置了。(或者可能还有另一个可接受的位置，但我们感到我们无法成功地把界线划在那里。)这一点与我在前面提到的原则能使个人摆脱即刻诱惑的功能相关。但在这里，需要我们克服的并非是即刻的诱惑而是即刻的**推理**。如果我到达了**那**点，我会推出：没有特殊的理由要我刚好在那时停下来，因而我最好在很早之前就停下来，那里有清楚的和一个**特殊的**界线。①

26 我想，正是这一点使原则能够来定义一个人。"**这些**是我划下的界线。"正是这些界线勾画或描绘了他，它们是这个人的外部边界。个人当身处非常幸运的环境时，知道他实际上不会向任何不可欲之地走很远，他可能就没有必要划下任何特殊的界线。因此，在这个意义上说，他可能没有身处不幸运环境中的人那么好界定。

象征效用

我们已经讲过：若采纳一项原则，则在这种情况下这一次做了特定的短视之行为，就意味着我们在未来会继续这样做。

---

① Thomas Schelling 的协调博弈理论也许会有用地纳入这个特殊性(specialness)概念。在试图与另一个人协作的过程中，我在寻找一种我们都认为是特殊的(但也是可欲的)行为，并且都认识到我们都认为它是特殊的——不仅是显著的(striking)，而且是特殊的。当有十个可选项时，其中九个是十分显著的，特殊的那个可能就是根本不显著的那个——至少在第一个层次上。

下面是一个协作问题。我们每一个人都要独立地在某一时期的德国哲学家之中挑出一位，而且如果我们俩挑出的是同一位，则我们每人都得到一份大奖。备选的哲学家有康德(Kant)、黑格尔(Hegel)、费希特(Fichte)、谢林(Schelling)和雅可比(Jacobi)。你会选择哪一个？

这一行为代表着为原则所不容的所有行为；做这个行为象征(symbolize)着做其余行为。我们前面讨论过，现在做某事与未来重复做某事之间有着两股纠缠不清的关联(现在做某事即会影响到你对自己再次做某事之概率的估计，也会改变你在未来做这件事的概率本身)。这个意指(meaning)、代表和象征的事实是它们构成的吗？或者象征是一个进一步的事实，并没有被上述两股关联所穷尽，而是其自身就会影响各种可选行动与结果的效用吗？我相信，象征是一股重大的进一步的关联，是一个恰当的决策理论所必须明确处理的。

弗洛伊德的理论，就是根据神经病患者的行为或症状的象征意义去解释它们的发生和持续的。尽管产生了显而易见的坏结果，但这些表面上不合理的行为及症状仍具有一种不那么明显的象征意义；它们象征着其他的某件事情，我们把它叫做 M。但是仅有这种象征意义并不能够独自来解释一个行为或症状的发生或持续。我们还必须补上：这些行为及症状所象征的东西本身——即 M——对该人具有某种效用或价值(或者，在要避免这些行为的情形下，指的是负效用或负价值)；不仅如此，这种被象征的 M 的效用又被归回给该行为或症状，因而赋予它的效用比看起来的要大。只有这样，才使行为的象征意义得以解释为什么要选择或展现该行为。弗洛伊德的理论必定认为，不仅仅行为及结果能够象征个人的进一步的事件，而且它们本身还能利用这些其他事件的情感意义(及效用价值)。有一种象征意义后，行为就会看作具有其象征指向之物的效用；一种精神症状的强度是会与其所代表之物相称的。[我并不知道在弗洛伊德的著作中是否有关于这个等式或一个较弱主张(即所象征之物的有些效用被归回到那个象征)的清楚陈述，即便我相信，在一些弗洛伊德式解释中已预设了这样的版本。若情感反应与某一

27

实际事件不相称，这可能标明此事件代表着与这些情感反应更相宜的其他事件或情况。][1]

要完成象征性行动，相对于行为者具有的其他选项而言，这个行为必须具有一个更高的效用，即一个代表了最大化目标(maximand)的更高数量。[2] 我已经表明过那是如何发生的。这个行为(或行为的后果之一)象征着某个情境，而这一被象征的情境的效用则通过象征联系被归回给该行为本身。请注意，标准的决策理论也认为伴随一种(概率性的)因果联系而来的效用归回。根据行为会确定地产生一种特定情境，该行为得以具有了——归因于它的——这种情境的效用；根据一个行为会概率性地产生某种情境，该行为得以具有了——归因于它的——该情境期望效用形式的效用。当前观点所增加的东西是：效用不仅仅可以通过因果联系，而且还可以通过象征联系回流(flow back)，即被归回。

在做行为决策中，起核心作用的不是表面上的因果联系(我这里思考的情形是行为人并不认为该行为自身是内在可欲的或有内在价值的)，而是行为与结果之间的象征性联系，一个标志就是在面对着强有力的证据表明该行为实际上并不具有所认定的那种因果后果时，我们还是坚持该行为。若有各种证据表明这种行为或策略具有有害的后果，有时候人们甚至拒绝去查证或甚至支持它们。(人们也许会以此为根据，主张强化毒品管制

---

① 一旦一个行为或结果开始象征其他的行为或结果，那么它的出现就会被看作其他行为或结果的证据，或者是作为它们的一个原因，但这是象征的结果而不是其原始构造的一部分(尽管这种证据性或因果性作用会强化这种象征的强度)。

② 最大化的决策理论就会这样假定。还有其他形式的规范决策理论，比如Herbert Simon的“满意”理论，但这也要求所做之行为具有一种超出(变动的)志向水平的效用，或者这种效用已经被归于它。

的措施**象征着**减少毒品的使用量，而最低工资法**象征着**帮助穷人。）一个改革者想要避免这些有害结果的话，他会发现有必要提出另一项政策，它能同样有效地象征迈向或实现该目标，且不具有那些坏的结果。简单地叫停目前的行动，则会从人们那里拿走它的象征效用，而他们是不愿意这样的事情发生的。

当然，把一个特定的象征意义赋予行为 A，这本身就具有某种因果后果，因为这影响到我们实施何种行为，一种纯粹的结果主义理论对此也会有话可讲。它会谈及赋予这种象征意义（或随后避免消灭这种象征意义）本身是否是因果上的最优行为。但这一观点并不同于一种行为 A 的纯粹后果主义（非象征）理论本身的观点，且它并不蕴含，我们必须只通过其因果后果来评价授予或宽容象征意义的那种行动。 28

由于象征行为常常是**表达性**(expressive)行为，它们的另一种观点是这样的：行为与情境之间的象征联系使得该行为能够表达某种态度、信念、价值、情感或任何东西。"回流"的并非效用，而是这种表达性(expressiveness)。沿着象征联系回流给这一行为的是表达某种特殊的态度、信念、价值和情感等（的可能性）。对于个人来讲，表达这点具有很高的效用，因而他会去做这一象征行为。①

个人为什么会选择做一个象征行为，两种构造方式对此的理解可能没有太大的区别。但对为什么**不**做一个象征行为，则两种方式均给出了不同的解释。第一种方式是效用沿着象征联系而归回给该行为，这就产生了一个困惑。很可能象征联系总

---

① 表达性也不是总是需要沿着象征联系回流。其他的东西也可以这样，并且这些东西对于本身就对行为者有高效用的行动产生出新的特征。关键在于回流的并不是效用。

是成立的，因此一个洗手行为总是象征着除掉罪恶之类的东西。因为这一被象征情境（即免除罪恶）大概总是具有很高的效用，如果效用被归回的话，那么洗手就总是具有最大的效用，那么人们为什么不总是去洗手呢？（很明显，这一情境的确发生在一些洗手强迫症的人身上，但不是发生在所有人身上，而且也不是所有的行为都是因为它们的象征意义才做的。）表达性理论说，作为一直存在的象征联系的结果，表达某种向往无罪的态度的可能性是一直存在的，但是表达出这点的效用却是随着语境而变化的，取决于个人表达有多久和有多么相关，也取决于个人的其他需要和欲望是什么，等等。表达那一态度或者感情的效用是要与其他的效用相竞争的。效用归因论对此将有不同的描述。被象征情境对于人们的绝对或相对效用会有所波动；摆脱内疚的效用能够在实际上变小，如果这个人最近采取了减轻内疚的措施的话——在现在（暂时地）就没有那么难处理了。或者尽管免除罪恶的效用仍然保持不变，但其他与之相竞争的物品（诸如吃饭）的效用暂时提高而变得比消除内疚的效用更高了。这两种用来理解象征表达性的结构都具有某种效用变动——其中有点细微的差别。我现在想强调的是：不管这种象征意义的确切构造是什么，它都是*重要的*。

当效用按照行为或结果的象征意义而被归于它时——也就是说，当行为或结果的效用等同于它所象征地意指之物的效用时——我们易于认为这是不合理的。当这种象征意义中含有童年时压抑的愿望与惧怕，或者某些当下的无意识的（unconscious）愿望与惧怕时，这很可能导致注定是挫败性的、令人不满意的和折磨性的行为。但是，基于无意识欲望的象征意义不是也有可能把这种满意反应添加到有意识地想要的东西上吗？无论如何，并非所有的象征意义都是根源于弗洛伊德式理论的。然而，

许多其他的象征意义,对于那一意义网络之外的人来说也是陌生的。想想有些人为了避免"丢脸"而承受的可怕后果,例如为了"维护尊严"而去决斗,为了"证明男子汉气概"去探险,这里所冒的是生命危险且有时候真为此丢了命。尽管如此,我们不应该太快地得出,压根不要任何象征意义或从来不依据象征意义来归因效用,会是一种更好的生活。

伦理原则规定(codify),对待他人的方式要适于他人的价值,适于我们对他们的同伴感情(fellow-feeling)。持有并遵循伦理原则,除了服务于这个具体目的之外,对我们还有一种象征意义。以尊重和回应的方式来对待他人(和一般意义上的价值),会把我们放在了那种价值的"同一边",也许把我们与那一边的每件事都结盟了,并象征着我们与此价值的交融。(它是比我们实际上介入我们的更大程度上象征这种交融呢?还是一种受欢迎的象征联系构成了实际的介入呢?)康德感到,当个人做出道德行动时,他是作为目的王国(kingdom of ends)的一员(即一个自由且理性的立法者)而行动。该道德行为并不能**致使**(cause)我们变成那个王国的一个(永久)成员,而是我们作为该王国一个成员会去做的,即它是我们在该环境下所做事情的一个例子,并因而象征着我们在该环境下的作为。这种道德行为和其他可能的事件与行动被归为一类,并得以代表和意指它们。由此,有道德(being ethical)得到了一种象征效用,它通约于其所代表的那些其他事情实际具有的效用。(那么,这要取决于这些实际上对人们所具有效用的进一步的东西——这是康德所厌恶要依赖的那种偶然性。)一个伦理行为对个人可以象征地意指各种各样的事情:成为一个自我立法的理性造物;成为目的王国中的一位立法成员;成为价值和个性的一类平等的根源和承认者;成为一个理性的、公正的、不自私的人;成为有同情心的

人；与自然和谐相处；回应有价值的事物；承认他人为上帝的造物。行为象征地表达和例示了这些伟大的东西，它们的效用被纳入了该行为的（象征）效用之中。因此，这些象征意义就成为个人去做道德行动的部分原因。有道德是象征（与）我们最重视之物（之间联系）的最有效手段之一。

30

我们生活的丰富性，有很大一部分在于象征意义以及它们的表达，即在于我们的文化和我们自己所赋予事物的象征意义。① 无论如何，若没有任何象征意义，也就是说我们欲望的任何重要性都不取决于这种象征意义，那很难讲我们的生活会是什么样子。那么我们想要什么呢？仅仅是物质的舒适、身体安全还有感官的快乐吗？我们在多大程度上想要这些东西，其中完全与它们对母爱和关怀的象征无关吗？只是为了财富和权力吗？我们在多大程度上想要这些东西，难道丝毫没有因为它们象征着从儿童的依赖中解放或终于可以成功地对抗自己的父母亲吗？而且也丝毫不是因为财富和权力所能带来之物的象征意义吗？仅仅是进化过程所灌输且植入我们的、天生无条件的强化物（reinforcer），而其他事物则只是作为有效的手段吗？这些曾帮助我们的祖先成为相关基因的更有效的创始者和保护者。我们应该选择这个作为我们唯一的目的吗？而且，如果我们高度重视它，难道我们不是也要重视象征着一种更有效的创始源头的任何因素吗？“不是的，若这与实际上的创始源相冲突，那么无论如何人们应当只重视实际上的繁衍或者保护后裔和亲

---

① 要注意，正如欲望与偏好一样，象征意义并不全部都是好的。关键在于一种合理性理论不一定要排除象征意义。然而，这些并不能保证有好的或可欲的内容。为此，人们又需要建立一种理论来辨别何种象征意义、何种偏好和欲望是可允许的，以此来约束何种特定的意义和欲望可以被纳入到更为形式化的合理性理论之中。

戚，以及进化过程所指明的对此有效的手段，也就是无条件的强化物和达致这些东西的手段。”（尽管如此，我们要注意，“为了最大化全面适合度”而灌输的那种欲望并不意指进化已经灌输了“要最大化全面适合度”的这种欲望。我想，男人们现在不会正在敲人工授精门诊室的门，要成为精子捐赠者，即便这样有助于增进他们的全面适合度。）但是，实际上导致的某个东西为什么比象征它要好得多，以至于象征就根本不值一提呢？“因为那就是底线，即实际发生的事情，其余的都是空谈。”但是为什么这条底线就比所有其他的线更好呢？

无论如何，如果我们是象征性生物的话——且人类学证实了这个特质之普遍性——那么这大概是进化使然。因此，象征化的诱人快乐，还有象征化的满足，也和其他的先天强化物一样有坚实的基础。也许象征能力有助于加强我们的其他欲望，且通过实际对象的强化来帮助这些欲望度过那种贫困时期。但是，无论进化解释是什么，这一能力就像其他的认识能力一样，并没有陷入其原初的适应功能。正如数学能力能够被用于探索抽象的数论和无穷论那样，尽管这种功能并不是进化所选择的，它也能够以其他有价值的方式被加以利用。一旦象征效用的能力存在了，它就可以使我们，例如，在某种意义上（即象征地）得到因果上或者概念上不可能的东西，因此从中获益，并且同时使我们有能力把与它们实际连接的特征分成好的与坏的，且经由只象征前者去获得它们。

31

这并没有否认象征意义和象征效用会有危险。象征意义很快就会涉及冲突，从而放大问题的意义，继而引发暴力。特别要避免与这样的情境有关的一种危险：一个行为尽管有很坏的因果后果，但因为它有极大的积极象征意义，我们依然还是会做这种行为。（回顾洗手强迫症和禁毒的例子。）理性的人会寻找另

一种不具有如此可怕的实际后果，但（几乎是）同样令人满意的象征行为。（但是，这并不意味着象征意义应该总是从属于并且词典式落后于因果产生的结果。）有时我们会认为一种象征联系比因果联系更好。如果一种后果（比如报复性伤人）是可欲的，但被认为是坏的，那么对此人来说，也许象征地达到这个目的就比施加实际的损害更好。[①] 如果我们能发现确立象征意义的那种连接类型的一般结构性准则，且这种准则能够区分好的与坏的象征意义，这将是很不错的。但也许我们必须要小心去隔离出某类情境（冲突是其中之一）和排除特定的象征意义。许多不可欲的象征意义一旦知道其起因后就不再平衡了，这是有用的；如果我们知道，什么产生了这些意义，它们在目前的行为中起了什么作用，那么我们可能不想再依据这些意义来做出这些行为。[②] 有些象征意义确实经受住了这些考验（例如，你为爱人所做的浪漫举动的象征意义）。也许关键在于要始终明白，意义和关联在何时是象征性的，要分别来追踪它们，并且不把它们（不知情地）作为因果上真实的来处理。这一点有助于许多弗洛伊德式象征意义，当作为象征而进入有意识的审思时，（如果它们被充分地"摸透"的话）它们就会失去其力量和影响。[③]

32

---

① 于是我们就应该区分开这样两种情形吗？一种情形是目标为 x 且人们以行动象征地去获得 x，另一种情形是目标为与 x 的象征性联系且人们以行动工具性地获得该目标。

② 关于在均衡中行动的一个讨论，见拙著 *Philosophical Explanations* (Cambridge, Mass.: Harvard Univ. Press, 1981), pp. 348–352。

③ 我感谢 Bernard Williams 提到了这个例子。当然，在缺乏一件事情被彻底摸透的独立准则时，括号里的最后一个从句排除了否证。Williams 还指出，有些象征意义涉及一个绝对不可能实现的幻想；而且也不清楚效用是如何能被赋予给不可能的情境的。然而，我并不想排除这一点，即甚至不融贯的情境对我们也可能有很高的效用。

象征意义也是具体的伦理决策中的一个成分。有人主张，不管是努力去救一个已知目前有危险的人(比如一个被困的矿工)，还是拒绝做出这样的努力都具有象征意义，这种意义还会影响我们的决策，即在现在去努力救人和预防意外措施上如何分配资源。(这一问题曾被称之为“与统计生命相对的实际生命”)。[①] 还有人主张，养活某人(维持生存)的象征意义进入了这样的讨论，即重症患者被容许以何种方式终结其生命——可以关闭他们的人工呼吸器，但不能停止食物饿死他们。[②] 我在《无政府、国家和乌托邦》一书中所提出的政治哲学，忽略了我们对社会纽带和关怀所做的共同的、正式的和严肃的象征陈述与表达所具有的重要性，因此(我曾经写过的)是不恰当的。[③]

我们生活于一个丰富的象征世界里，部分程度上是文化上的，部分程度上是我们自己的个人创造，有了情境所具有的意义，我们也因此避开了或扩展了我们情境的限制，不仅是通过幻想，而且也体现在行动中。我们归于行为和事件的效用是与其象征之物相协调的，而且就像对待它们所代表之物那样，我们力争去实现(或避免)它们。[④] 这样就需要一种更宽泛的决策理论，能纳入这些象征联系，并且厘清这些东西所引入的新结构。

---

① 参见 Charles Fried, *An Anatomy of Values* (Cambridge, Mass.: Harvard Univ. Press, 1970), pp. 207 - 218。

② 参见 Ronald Carson, “The Symbolic Significance of Giving to Eat and Drink”, in *By No Extraordinary Means: The Choice of Forgo Life-Sustaining Food and Water*, ed. Joanne Lynn (Bloomington: Indiana Univ. Press, 1986), pp. 84 - 88。

③ 参见我的 *The Examined Life*, pp. 286 - 292。

④ 对于有些商品广告如何运用了这一现象的讨论，请见拙著 *The Examined Life*, pp. 121 - 122。

在社会科学家中，人类学家最为关注行为、宗教仪式、文化形式和文化实践等的象征意义，还有这些东西对一个群体当前生活的重要性。① 他们的工作就是要阐述清楚，所以只引入一个相对粗糙且含混的象征意义这一概念的话，那就多少有点尴尬了。尽管如此，这个概念还是有其用处，只是与形式结构不容易相关联的那些细致讨论就无法服务于此了。若行为的象征意义（即其象征效用）能被纳入一种（规范的）决策理论，我们可以据此把合理选择理论与人类学关注更紧密地连接起来。这种连接可以具有两个方向。第一个是上升的方向，依据纳入象征效用的个人选择行为来解释社会模式和结构。我在这里所建议的方向并不是这一方法论个人主义和还原主义的方向。② 第二个是下降的方向，解释人类学家所描绘的社会意义模式如何对个人的行为产生影响：经由个人在决策中赋予象征效用以权重。33 （一些人类学家，出于职业性傲慢，似乎并不关心他们所勾画的文化意义与个人行为是如何协调的。）

行为（或结果）的象征效用是如何起作用的呢？这种象征的联系或联系链的性质又是什么呢？效用或表达性的可能性是以何种方式沿着这一链条从所象征的情境流到这一象征行为（或

① 参见 Raymond Firth, *Symbols: Public and Private* (Ithaca: Cornell Univ. Press, 1973); Clifford Geertz, "Deep Play: Notes on the Balinese Cock-fight", 载于他的 *The Interpretation of Cultures* (New York: Basic Books, 1973)。

② 实际上，就象征意义是被社会地创造、维续和协调的而言，正如它也受社会因素所限制一样，我们在此也许发现了方法论个人主义之解释的一个限制——给定这种意义的影响和后果，这是一个重大的限制。因为一种象征效用可以是社会性的，不仅仅是受到社会的塑造和共享（也就是说，对于社会中的许多人来说是一样的），而且还在于被看作是共享的（该特质是内在于它具有那一象征效用之中的）。因此，方法论个人主义解释如何能够应对所包含的这种复杂性，这是不清楚的。无论如何，暂且不论象征意义，一种语言的方法论个人主义论说会是什么样子，这也是不清楚的。

结果)的呢？我们首先要注意到，象征意义不止是采纳一个原则而使得某些行为得以代表其他行为的那种方法。那里，一个行为代表的是同一类之中的其他事物——其他的行为——或者代表了整类事物，而象征意义能够把一个行为与不是(一类)行为的事物联系起来，例如，和做某种人或实现某种事态联系起来。

古德曼(Nelson Goodman)为我们提供了一些有用的和有启发性的分类。[①] 依据古德曼，当 *A* 指示(refers to) *B* 时 *A* **指称**(denote)*B*；当 *A* 指示 *P* 且 *A* 是 *P* 的一个例子时，*A* **例示**(exemplify)*P*，也就是说为 *P* 所指称(要么字面上要么隐喻地)；当 *A* 指示 *P* 且 *A* 比喻地或隐喻地具有属性 *P*(因此 *P* 比喻地指称 *A*)时，*A* **表达**(express)*P*，而且，在例示 *P* 时，*A* 作为美学符号(aesthetic symbol)起作用。这些关系能形成链条。当 *A* 指称某个 *C* 且这个 *C* 例示 *B* 时，或者当 *A* 例示某个 *C* 并且 *C* 指称 *B* 时，*A* **暗示**(alludes)*B*。更长的链条也是可能的，[②] 而有些关联将是比喻的或隐喻的。这些链条和其他链条能够把行为与更进一步的更大的情境或状况关联起来，后者是该行为能够象征地代表(represent)或影射的(等等)，然后这些更大情境的效用就为该行为自身提供一种会进入有关它的决策的**象征效用**。这些链条不一定很长：若 *A* 是术语 *P* 的字面意义上的外延，*B* 是那个术语的隐喻外延，则 *A* 可以让 *B* 作为其象征意义的一部分。有时一个行为可以作为某物最好的具体实现(我

---

① Nelson Goodman, *Languages of Art* (Indianapolis: Bobbs-Merrill, 1968), pp. 45 - 95. 对于 Goodman 的美学价值论说的讨论，参见我的“Goodman, Nelson on Merit, Aesthetic”, *Journal of Philosophy* 69 (1972): 783 - 785。

② Catherine Elgin, *With Reference to Reference* (Indiannapolis: Hackett, 1983), p. 143, 讨论了具有五个连结的特定链条。

们能做得最好的)而象征地意指该物。①

经链条而与行动联系起来的那个更大情境的效用,还有这个链条本身的性质,会以何种特定的方式来决定该行为的象征效用(或表达性)呢？是否链条越短,更大的情境传递到该行为本身的效用(或表达性)就越多呢？存在的关联越多,效用/表达是否就丢失得越多呢？关联的类型不同,更大情境的效用(或表达的可能性)的传递比例会否不同呢？(我认为,行为的象征效用不会比链条所联系的更大情境的效用更大,但可以比它小。)

34 只有某些象征联系才会引起效用的归回吗？决定象征联系类型的又是什么因素呢？这些都是选择在确定性情境之下出现的问题;选择在风险和不确定性下还会引发进一步的问题。沿着某些特殊的链条是否会存在一种概率式折扣呢？有些类型的更大情境,即使不是确定会发生,依然能把它们的全部效用都传递给可能产生它们的那一行为吗？当然,一个行为伴有特定的风险或不确定性,这个事实本身也许会赋予该行为一种特定的象征意义和效用,也许与一个勇敢和有胆量的人相联系,或者与一个莽夫相联系。尽管有时候,没有确定性,而是出现概率会完全消除一种象征意义。但情况也不是这样的,即有50%或10%概率实现某个目标本身总是具有那一目标的50%或10%的象征效用——它甚至不一定象征那一目标,即使是在部分程度上。象征效用为什么必须被视为决策理论的一个独立的部分,而不是简单地纳入现有的(因果的和证据性的)决策理论,这里有另外一个理由。因为这种象征效用并不遵守一种期望价值公式。观

① 行为的象征效用能被看作是对该行为的一种解释,即以某种方式来看待自己或该行为的一种方法,由此诠释联结的各种模式以及诠释自身的诸种全部理论能够进入对象征效用的规定之中吗？

察到的期望值公式和相关决策理论的公理出现某种偏差，我们可以尝试将此归于象征效用的出现去理解和解释。我心中想到的是阿莱悖论（Allais Paradox）、确定性效应、违背萨维奇确信原则（Savage's Sure Thing Principle）等等。确定性本身对于我们具有一种象征效用。90％与 100％之间的概率差别，其影响远远大于 80％与 90％之间的概率差别，尽管若把这些差别都嵌入到更大的、除此之外就是完全相同的赌博当中去，这些差别之间的差别就会消失——这种消失标明这种差别是象征的。[①] 象征效用的详尽理论还有待发展，我们现在所能做的，就是在一种更一般的决策理论架构内标出它的一席之地。我在下一章中对 35

---

① 或者不过是有些数字本身就更为突出，而其效用受到了这种突出性的影响？两位数的膨胀的象征意义是膨胀失控了，所以相对于从 16％到 17％而言，我们更关注从 9％ 到 10％的膨胀变化；如果我们以 11 为基础开始计算，那么（象征）线就要定在其他地方。在 *Anarchy, State and Utopia* 一书中，我讨论过完全排除掉一个问题的意义，因此邪恶场景从 1 减少到 0 比从 2 减少到 1 的差别更大。我曾将此作为一个空想家（ideologue）的标识（p. 226），但它最好是被看作象征意义的一个标识。

要注意，当确定性效果出现时，效用要求由一种有点不同于通常程序的程序来加以测度。按照通常的程序，两种结果 x 和 z 的效用量是依据对它们两者的偏好次序而分配的，并且任何第三个事物 y 的效用是依据阿基米得条件（Achimedean condition）而得来的。该条件是说当 x 偏好于 y，且 y 偏好于 z，那么就有唯一的概率 p（在 0 和 1 之间）能够使得确定的 y 无差别于这个选项，即由概率为 p 的 x 和概率为（$1-p$）的 z 所组成的选项。当人们满足所有的冯·诺伊曼-摩根斯坦（Von Neumann-Morgenstern）条件时，那么就不存在任何问题；但是当确定性效应出现时，那个确定的居间项 y 会得到一个错误的效用。一种更好的程序也许是不考虑任何确定后果的情况下来衡量效用，把前面处理的东西嵌入到经典概率混合物（cannonical probability mixture）中，比如说概率 1/2。然后人们将被要求找到一个概率 p，以致下面两项对他是无差别的：即一项是一半概率得不到任何东西和一半概率得到 y；另一项是一半概率得不到东西和一半的概率得到概率为 p 的 x 和概率为（$1-p$）的 z。我们由此控制了确定性效果。当然，只有当这样的程序在经典概率混合物中对特定的概率（这里是 50％）不敏感时，此程序才起作用。情形将可能是在经典概率混合物中，在一个大范围的概率内，也许是所有除了 0 和 1 这两个极端点之外的所有概率，都要得到相同的结果。

此会有更多的论述。

目的论装置

原则把证据性支持或概率从一个情形传递到另一个情形，从而有助你发现真理。原则也把效用从一些行动传递到另一些行动，以这种方式帮助你克服诱惑。原则既是概率传递装置，也是效用的传递装置。①

原则具有各种各样的功能和效果：智识上的、内省的、个人性的和人际间的。但这并不是说，在每种可能的情境下，原则都会有这些效果。一种气温调节机制只在某一气温范围内才能工作；超出了那个范围，它就不能把温度调回来，甚至会融化或者结冰，这取决于它的材料。为什么进化没有给我们更好的体温调节机制呢？既然这种极端情形发生的概率很小，那么在能量和附带的其他功能牺牲上的代价就太高了。即使一种机制在某些可能出现的情境下不起作用，但它还是能够相当好地，即足够

36

① 所有的原则都必定只能传递两者之一，还是有的原则能同时传递两者呢？我们可否推测，有一种东西，即概率和效用 $p_i \times u^i$ 是所有原则都能够传递的呢？在决策理论内部没有单独的术语来指称这种加权和，$p_i \times u^i$，尽管它们经常作为一个单位(unit)一起出现。实际上，形式理论必须置入非常特殊的程序来分开它们，这种程序频繁地假定在特定的情境下已经把它们成功地分开，之后就运用工具把这点扩展到一般的情境。通过把概率和效用视为一个整合数量——称之为重要性——而不是太快地把这些成分分开，即通过探究这一整合数量满足什么样的条件，我们可以学到一些有益的东西。[但是，由于重要性能被嵌入到概率混合物之中，那么一开始在两种组成部分之间是否就有一种不对称呢？我们需要探究相对应的效用混合物的可能性(它可能会扩大或者缩小组成部分的重要性)吗？时间因素是否在开始时就包括在组合之中，只是后来才被抽离出来作为一个成分？时间偏好主要是一个有关概率和效用的问题，还是时间距离本身就构成了重要性的降低？效用在时间上的延伸——不是在时间上的转移——扩大了它的重要性吗？]

好地发挥其功能。这对原则来说也是一样的。

为了证成一项原则，你规定了其功能并表明，在给定的成本、约束等之下，它可以有效地履行该功能，还比其他原则更为有效。我们也可以质疑该功能的可欲性。为什么要**有某个东西**有那种功能呢？一种证成将表明（或认为）：该功能是可欲的，还不影响其他更可欲的功能。完全确定后，一项原则 P 的证成就具有一种决策理论的结构，原则 P 具有一个行为的地位，从而可与特定的其他选项相竞争，具有某种概率来达到具有某种可欲性的目标等。（原则的一种最优选择理论要考虑各种因素，我们之前对这些因素的讨论契合于这个决策理论框架，即目的论框架。）

原则的设计可以是用来应对特定的情境或预防**特殊的**危险，诸如屈服于即刻的诱惑、偏袒于自己的利益和把自己所希望的东西信以为真等。因此，不面对那些危险的人并无必要拥有**那些**原则。而且除了原则以外，可能还存在其他装置能对付这些危险。（个人不仅仅可以通过原则，他还可以通过与他人的移情互动，或想象完全投入他人情境，从而避免偏袒自己的利益。）

我们也许会追问，一般性原则这一装置是否有其**自身的**偏见或者缺陷。根据决策理论来看待事物，这使得我们能把原则看作（被假定）具有特定效果（即它们的功能）的装置，因此不仅可将一些原则与其他原则相比较，还可把原则与其他装置相比较。有些目标是原则不可能或极难达到的，然而其他的手段也许能更容易地达到**那些**目标。

如果我们的一个重大目标就是要生活在一起，不具有会撕裂或毁掉可贵社会制度的那种紧张冲突，那么当冲突双方都强硬地提出了互不相容的原则时，它们也许就没有办法让双方都同意任何一个第三原则，更不用说原来的那两个原则，来解决那

个冲突了。也许需要的是某种妥协——但妥协正是原则不该做的事！因此一个制度或国家的领导者就是要让事情继续下去，37 做出某些安排，平息人们的怒火，维续制度生活。的确，可能会存在某个原则推荐我们这样做，它适用于每种原则性冲突，即使这种冲突有非常严重的威胁，甚至有可能撕裂可贵制度或使之瘫痪。然而，妥协的内容可能不过是由竞争力量(force)[给定其各自的能力(power)]能成功忍受的东西所决定的。因此妥协自身没有必要在这种意义上由原则所决定，即其细节要看作是为其他类似情境设定了先例。

这并不是建议政治和制度的领导者要无原则。也许除了罕见的情况之外，他们在决策和行动中是有原则的，在罕见情况下，上述(above-stated)原则命令(无原则的)妥协要生效。(然而，看看美国政府的结构，其中似乎存在着一种不同的分工：有些类型的决策——那些由司法部门做出的——被认为是要求有原则的，然而其他决策的细节——那些由最高执行官和立法机构所做出的——一般来说是留给各种力量进行角逐的，同时也受到司法部门的某种监视以免违背某种一般性原则。)我在这里打算指出的唯一一点是：原则这样一种目的性装置并不是适合于每一种目的的。

那么，这是否意味着我们必须应用某种原则来决定何时援用一项原则呢？什么因素使得某一情境要受原则而不是其他东西的指导呢？任何事情都应该基于原则而决断，这种观点可能很乐意把那也作为一项原则而提出来。然而，认为这些而不是那些事情应基于原则来决定，这种观点也是一种有原则的观点吗？它必定是吗？若在某种情境下，偏爱其他模式的特定理由也能解决问题，那么是否存在一种支持依原则来决定事情的前设(presumption)呢？如果不存在，那么以原则为基础来决定某

件具体的事情不是会产生这个问题，亦即，**那时**为什么要诉诸原则呢？这样做是否会使个别的选择偏离其他模式本会产生的那种结果呢？或者是不存在任何前设，而仅仅存在着一种决策理论论说，它决定什么时候原则是适宜的。而且这种论说本身也使用了某种决策理论原则，并且还预设了一点，即至少对于这一情形确定何时原则是适宜的，则使用(决策理论中的)某个原则就是适宜的。

行为原则具有一种目的性功能，这种看法的另外一个理由是这样的。一个实际的例子，比如说纳粹德国，就可以彻底破坏或驳倒将会赞同或允许纳粹德国的原则 P。但是为什么假设例子就不够呢？在 1911 年人们难道不能说：原则 P 会允许(甚或在某些情势下会要求)纳粹德国(之类的东西)，因此，P 是错误的、不可接受的、邪恶的吗？

38

然而，如果原则只是被认为涵盖了将(will)、本会(would)以及能够(could)出现的情况，那么在事实发生之前，如果它被认为是一个不可能出现的情形(即情境、动因或会导致它的任何不能出现或成功的东西)，那么也许不管是这个原则还是任何别的原则，它都不会被认为是一个**相关**的反例。但是一旦人们发现人性**能够**做到那点——因为做过——则赞同它的原则 P 就被驳倒了。

人们接受并依一项原则而行动，由此产生的后果有可能会贬低该原则。“他们按原则 P 行事，看看现在引起的可怕情境吧。”也许其他人会说这是因为他们跟随原则走得太远，或者走的是错误的方向，也就是原则自身并没有**要求**他们这样做。尽管如此，原则 P 仍然被贬低了。当每个人都对遵循 P 的早前结果心怀厌恶的时候，人们很难说：“让我们再次遵循 P，不过这次以正确的方式吧。”为什么呢？因为即使 P 并没有要求那样做，

但 P 本身不是适宜于那种行为方式吗？这就是当人们遵循 P，正如通常所做的那样时，接受原则 P 所带来的后果。[①] 如果原则是为了获得某种效果的装置，那么它就是**当得到遵守时**而具有的那些效果的装置；因此，当我们把原则作为一种目的性装置进行评估时，相关的就是遵守它时实际发生的事情，而不仅是它**宣称**的事情。

但原则不也同时是基本的事实(basic truths)吗？通过统摄事例(在 Hepel 那里)，它有助于我们理解并由此解释它们为什么会成立吗？此处，原则也可能再次被认为是一种装置，它具有产生理解的**认识论**功能。这样一来，即使是在这里，我们也可以(在决策理论上)询问：是否还有其他的理解路径，是否这些路径更适合于某些语境或对象等？

然而，原则可能是使得特定的真理为真的因素吗？也就是说原则产生了这些真理，在这种情形下，原则的首要性是**本体论的**。如果这不仅仅是重复其认知功能——我们通过原则能最好地理解特定的真理(truths)——而且如果“产生”不是一种时间关系，“使之为真”也不是一种因果关系，那么就不清楚这个本体论观点究竟是什么主张。无论如何，我们不需否认，不管现象与原则之间的关系是本体论的、认识论的，抑或是混合的，构建(数学的、自然现象的、心理学方面的)原则都可以使得现象之间更为融贯，我们的理解也会更为深入。因而，原则除了具有我们开始讨论的那种智识功能，即传递支持与概率之外，还有一种更深入的智识功能，即深化、整合并明确我们对于原则所关涉对象的

---

① 在我的 *The Examined Life* 中，见“理想与现实”那一节。这里同时开放出了这样一种可能性：那些不想使原则 P 得到遵循而产生出某一结果的人们，可能做安排使 P 得到遵守，从而导致一种更坏的可怕结果，由此贬低 P。

理解。(这一功能在支持和概率之间产生了更紧密的关系。这些是*构成*了而不是来源于增强的理解吗?)因此,道德原则的构建能够深化我们对于道德行为、道德事实与道德现象的理解。但在这里,道德原则具有一种无异于物理原则或心理原则的地位,它们描述现象但没有显而易见的理由要*根据它们*行事。人们也许会说,尽管正确的道德原则是成立的——所以它们*应该*被遵循——但是实现它们的唯一方式(也就是符合我们的实际行为)就是要努力遵循它们,即*按照它们*行事。这是一种经验主张,是需要证据的主张。或许我们能够构建并遵守的原则远远达不到正确的道德原则——更复杂的道德真理——所要求的,以至于我们只有通过不遵循依原则行事的路径才能更好地符合后者。归根结底,这是一个经验性的问题。无论如何,这又一次使得依原则行事成为一种目的性装置。 39

*原则*这一术语通常用来指向比规则更深和更一般的东西。原则是细节于其中能找到位置的大纲。在契约的协商中,这些大纲可作为指南:各方首先是就管辖国家之间的和平条约、公司之间的合并和建立新学校等的原则取得一致,随后才去敲定细节。如果我们在一般原则框架上达成一致的话,那么这能够在随后节省时间。如果细节的选择必须在契合于这些一般原则的选项中进行的话,那么就不必讨论和辩论每一种可能性,不是每一个问题都要重新开始,获得同意的原则本身就可以用来解决争端。然而,原则的一般性并不限于指导行为。教科书的标题宣称它们提出的是经济学、物理学或心理学诸原则,亦即是这些科目的细节能于其中找到其从属位置的总体陈述框架。或许这里也有指导作用:所提出的原则将指导读者们如何理解各个科目的细节。

如果我们的认知能力受到某种方式的限制,那么以下两者

就可能有所差别：一是能够指导我们理解某一科目的原则；二是支撑该科目的所有事实的一般性的真陈述(或法则)。后种一般性可能会太过复杂，以致我们无法理解它，根据它派生出后果，或者在其中镶嵌细节。那么，哪一个会是(比如说)物理学原则：是我们能够理解、能够更好地符合的那种大致准确的一般性，还是我们无法理解也无法操作但确切为真且涵盖一切的陈述呢？

康德式传统倾向于认为，这种原则的功能是要指导自觉反思性造物的审思与行动；因此，原则既有一种理论功能，也有一种实践功能。我们不是一种不经任何指导就自动地行为的造物。我们能够想象拥有自动的指导——这不是使得原则对我们来说成为多余的吗？——或者，更切题的是，以一种无需任何指引的方式来行动，比如说，随机行为。[完全随机的行为足以使我们摆脱因果性领域(这是康德为原则保留的功能)吗？]这不是表明原则的目的就在于指引我们得到随机行为无法让我们得到的某些东西(无论那是什么)吗？这不是把原则作为目的性装置吗？然而，康德还坚持原则是对我们的理性本质的表达，即合理性的构成成分。理性地思考或行为就是要符合(某类)原则。因此只看待原则所服务的**外在**功能就是错误的。如果原则是只有一个理性行为者才能构建与运用的东西，而且如果成为理性的是我们所重视的，那么遵守原则就能够象征且表达我们的理性。原则因此就对我们具有了很高的效用，不是因为使用它们而可以获得的东西，而是因为它们的使用所象征和表达的东西。就此而言，原则并不仅仅是目的性的装置。但这里仍然还有一个问题，如果我们的理性本质并不服务于任何更进一步的目标，那么我们为什么要如此珍视我们的理性本质，以及如此珍视表达这一点的依原则和理由行事呢？为什么要就此止步呢？

原则与理性为什么是如此密切相关呢？我们为什么要重视理性呢？把一件事情、一个行为或一个信念称为合理的，就是评估做它或持有它的理由（以及人们考虑反对做它或相信它的诸理由的方式。）。如果理由据其本性是一般性的，且如果原则把握了出于这种一般性理由而行动的观念——这样人们就承诺了自己在其他类似的情况下也会这样做——那么要合理地行为和思考，个人就必须按照原则来思考与行动。但是我们为什么应该合理地思考和行为呢？一种回答可能是，我们是理性的，我们具有合理地行为的能力，而且我们重视自己的这个样子。① 但是如果不想止步于这种自大，我们不是必须诉诸合理地相信和行动所服务的诸种功能吗？为什么理由必须是一般性的呢？把一般性的理由与最类似于它们的非一般性近亲做一比较。要解释为什么应当使用理由而不是这些替代项的话，我们就必须再次援引理由的诸种功能。

因此，我们的问题由一个关于原则的问题而转变成了一个关于合理性的问题。合理性是为了什么？合理性的功能是什么？合理性自身是完全目的论的，即完全工具性的吗？这些问题推动了以下章节。

---

① 当我们的行为是有意识地由一种理性法则所决定，且这种理性法则是构成我们本质属性的一个原则时，我们就是自由的。对这一观点的一些批判思考请参见我的 *Philosophical Explanations*, pp. 353 - 355。

2. 决策价值

经济学家和统计学家已经发展了一种精致的合理决策理论,并在理论研究与政策研究中加以广泛应用。这种理论很强大,数学上精确,易于操作。尽管它描述实际行为的恰当性广受质疑,但它依然是对于一种合理决策应满足什么条件的支配性观点:它是一种支配性的规范观点。我相信标准的决策理论需要加以扩展,从而明确地纳入行动的象征意义等要素。纽科姆难题(Newcomb's Problem)为我们探讨现有标准理论的不当之处提供了一个很有用的切入点。我的目标是构造出一种宽泛决策论,以恰当地处理和涵盖这一问题,然后再应用这一宽泛理论来阐明囚徒困境(Prisoner's Dilemma)。后者准确地明确(sharpen)了合理的社会合作问题,以及为了维持合作有时是否有必要采用强制,由此在最近极大地推进了形式(formal)社会理论。

纽科姆难题

纽科姆难题众所周知,故我在此只做简要的描述。① 有个

---

① 这个问题是由一位物理学家 William Newcomb 想出来的,我是从我们俩共同的一位朋友那里得知这个问题的,而且这个问题(经过 Newcomb 的同意)是我首先拿出来讨论的,参见 Robert Nozick, "Newcomb's Problem and Two Principles of Choice",载于 *Essays in Honor of C. G. Hempel*, ed. N. Rescher et. Al. (Dordrecht: Reidl, 1969), pp. 114 – 146。

存在物，你非常确信它有能力准确地预测你的选择，会预测你在下述情境中的选择。有两个盒子，*B*1 和 *B*2。*B*1 盒子里面装有一千美元；*B*2 盒子里面或者装有一百万美元（*M* 元），或者什么都没有。你要选择下面两个行为之一：(1) 拿两个盒子里的东西；(2) 只拿第二个盒子里的东西。进而，你知道，而且那个存在物也知道你知道，如此等等，如果该存在物预测到你会拿两个盒子里的东西，那么它就不会在第二个盒子里面放进 *M* 元；如果它预测到你将只选择第二个盒子里的东西，那么他就确实会在第二个盒子中放进去 *M* 元。该存在物首先做出它的预测；然后根据它自己的预测，决定是否把 *M* 元放在第二个盒子里面；最后你再做出选择。

42 这个问题不仅仅在于你要决定怎样做，而且也在于你要准确理解：两种相冲突但强有力的论证之中，哪一个出了错。第一种论证是这样的：如果你选择要两个盒子里的东西，该存在物几乎肯定会预测到，并且不会在第二个盒子里放进 *M* 元，因而你也几乎可以肯定地只得到一千元；而如果你选择只拿第二个盒子里的东西，该存在物几乎肯定可以预测到这一点，并会在第二个盒子里放进 *M* 元，因而你也几乎肯定可以得到 *M* 元。由此，你应该只拿第二个盒子里面的东西。第二种论证则是这样的：那个存在物已经做完了它的预测，并且第二个盒子要么已经是放了 *M* 元，要么是没放。此时第二个盒子是否有 *M* 元，这已经是不变和确定了的。如果该物在第二个盒子中放了 *M* 元，那么你拿两个盒里的东西，则你将得到 *M* 元外加一千元；而如果你只拿第二个盒子的东西，那么你将只能得到 *M* 元。如果该存在物在第二个盒子中没有放 *M* 元，那么如果你拿两个盒里的东西，你会得到一千元；而如果你只拿第二个盒子里的东西，你这次就一分钱也拿不到了。因此无论在何种情况下，即无论

是否放了 M 元,你拿两个盒子的东西都会多一千元钱。[这也就是说,拿两个盒里的东西就是所谓的"**占优于**(dominates)"只拿第二个盒里的东西。] 因此,你应该拿两个盒子里的东西。

我在 1969 年首先引入并讨论了这个问题,自此,对这个问题就有了更详尽的讨论和做了富有阐示力的理论化。① 在我的初始论文中,我区分了两种条件概率:一种条件概率标出了一个行为能**影响**获得何种后果,还有一种并不标出这种影响。我指出,当一行为与占优原则相冲突时,若该行动的条件概率属于第二(无影响、无作用)类,那么就不应该援用条件期望效用最大化原则。我是通过直观性例子来支持它的。[因为试图纳入某种自反性(reflexivity),这些例子比起随后其他人所讨论的例子要多少复杂一些。]我主张,即使患某种疾病与选择某种职业之间存在着某种基因倾向关联,这也不会导致一个人因为会提高她患上这种病的概率估计而避免选择这一职业——因为实际上无论她是得到了基因修复,还是得了病,这都不会受到职业选择的**影响**。我过去根本没想到要用这个论点来充分且系统地发展相竞争的各种版本的决策理论。这些理论可能是证据性的,也可能是因果性的,并且随之对期望效用原则甚至是占优原则也有不同的版本。②

---

① 截止到 1985 年的一个论文选集,还含有其他人的书单,请见 *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem*, ed. Richmond Campbell and Lanning Sowden (Vancouver: University of British Columbia Press, 1985)。

② 对于因果决策理论,参见 Allan Gibbard and William Harper, "Counterfactuals and Two Kinds of Expected Utility", 载于 *Foundations and Applications of Decision Theory*, ed. C. A. Hooker et al. (Dordrecht: Reidel, 1978), 重刊于 *Paradoxes of Rationality and Cooperation*, ed. Campbell and Sowden; David Lewis, "Causal Decision Theory", *Australasian Journal of Philosophy* 59 (1981): 5-30; J. H. Sobel, (转下页)

43 最大化期望效用这一传统原则把行为 A 的期望效用($EU(A)$)视为其可能的(互斥的)各种结果所产生效用的加权和,权数为其概率,其和为1。

$$EU(A) = prob(O_1) \times u(O_1) + prob(O_2) \times u(O_2) + K + prob(O_n) \times u(O_n) = \sum_{i=1}^{n} prob(O_i) \times u(O_i)$$

一个更恰当的原则会注意到后果并不一定在概率上独立于行为,它规定期望效用不是由后果的简单概率加权的,而是既定行动的结果条件概率加权的。把这个称为 A 的证据性期望效用,表示为 $EEU(A)$。①

$$EEU(A) = prob(O_1/A) \times u(O_1) + prob(O_2/A) \times$$

(接上页)"Circumstance and Dominance in a Causal Decision Theory", *Synthese* 63 (1985)。

我还没注意到这种特定情境的可能性,那里状态在概率上独立于行为,然而受到它们的因果影响——Gibbard 和 Harper 的 Reoboam 例子——这本应该在我初始论文的第132页的三行图标中标出一个第四行。

① 对最大化条件期望效用的讨论,尽管没用**证据效用**(evidential utility)的术语,参见我1963年 Princeton University 的博士论文,*The Normative Theory of Individual Choice* (rpt. New York: Garland Press 1990),参见232页:"要用来决定一个行为的期望效用的概率,现在必须是在给定行为已完成的情况下各种状态的条件概率。[一般而言这是正确的。然而,当状态的概率独立于行为时,则在给定某一行为已完成的情况之下的每个状态的条件概率就将等同于该状态的概率,这样就可以使用后一种概率。"]这里陈述的条件期望效用的公式仍然是针对那时所探讨两种具体行为的情形的,尽管不是针对各种可变行为的普遍公式。普遍公式呈现于 Richard Jeffrey, *The Logic of Decision* (New York: McGraw-Hill, 1965)

我们在本书中所关注的所有问题都是出现在概率明确界定的情况下,无论这个概率是有条件的还是无条件的,无论是主观的还是客观的。其他的问题使得一些论者利用概率区段(intervals)来构建理论;这方面的例子,请见 Issac Levi, *Hard Choices* (Cambridge: Cambridge Univ. Press, 1986)。这里所表述的观点可以被多么精确地在这种框架之中得到重述,这是一个有待探究的问题。

$$u(O_2) + K + prob(O_n/A) \times u(O_n)$$
$$= \sum_{i=1}^{n} prob(O_i/A) \times u(O_i)$$

因果决策理论家同样不仅使用结果的无条件概率，还使用后果与行为相关的概率，这次不是简单地全用条件概率 $prob(O_i/A)$，而是直接指明因果影响的某种因果概率关系。带有这些因果概率的对应公式表示了行为 A 的因果期望效用，记为 $CEU(A)$。

尽管有这些和其他的技术化阐明——回溯假设(backtracking subjunctive)、明确地纳入提醒与元提醒(tickles and meta-tickles)、决策的可核准性(ratifiability of decisions)，等等——尽管有人竭力表明该问题的界定错误或不融贯是无可救药的，[①]但争议丝毫没有缓解。没有一个办法具有完全的说服力。

纽科姆问题是个复杂的问题。其他情形还会包含更进一步的复杂性，各方的推理过程似乎都相当有说服力——且我们是可错的造物。对于这种情形而言，绝对信任任何一种特定的推理思路或决策原则都是没道理的。[②]

---

① 试图把这一问题作为错误地形成的、错误地界定的或在原则上是不可能的而拒绝，包括 Issac Levi, "Newcomb's Many Problems", *Theory and Decision* 6 (1975): 161 - 175; J. L. Mackie, "Newcomb's Paradox and the Direction of Causation", *Canadian Journal of Philosophy* 7 (1977): 213 - 225; and William Talbott, "Standard and Non-standard Newcomb Problems", *Synthese* 70 (1987): 415 - 458。针对许多这类批评而对这个问题所做的一个辩护，参见 Jordan Howard Sobel, "Newcomblike Problems", *Midwest Studies in Philosophy* 15 (1990): 224 - 255。

② 几年前在研究生的研讨课上，有几个学生，特别是 David Cope 提问道：给定两边都有强大的论证，人们如何能够确定是因果的决策理论还是证据的决策理论呢？我很感激这次讨论，因为它开启了我接下来继续思考的轨道。[但是 Howard Sobel 写信对我说，事情并不是对称的，因为只有因果论者尝试为他们自己的立场提供论证且去诊断反方论证之中的(所谓)错误，这与我在初始论文中所提出的要件(desiderata)是一致的。]

44 第一个盒子中的钱数(一千美元)很少为人所注意。[①] 如果占优论证——即上述第二种观点——是正确的,那么你拿两个盒子里的东西会更好,即便第一个盒子当中的钱要少得多,比如说只有一美元,甚至一便士或万分之一概率得到一便士。然而,在这种情形下,极少有人会选两个盒子,而承认另一种观点——即如果我们只拿第二个盒子里的东西,我们将几乎肯定可以拿到 M 美元——不具任何力量。另一方面,如果以上第一种观点是正确的并被理解为一种期望效用论证(嵌入的条件概率并不需要表达任何影响),那么尽管第一个盒子里的钱数 X 可以变得远远多于一千美元,但是这个人仍将选择只拿第二个盒子里的东西。让我们假定成功地预测你的行为(对于你每次可能做出的选择)的概率是 99%,u 表示效用函数,只拿第二个盒子里东西的期望效用就是 $0.99u(\$M)$,然而拿两个盒里的东西的期望效用就是 $0.99u(X)+0.01u(\$M+X)$。如果我们假定钱的效用在这个范围内是与其数量呈线性关系,那么选择两个盒子的期望效用就是 $u(X)+0.01u(\$M)$。这种情况下,如果 $0.99u(\$M)$ 大于 $u(X)+0.01u(\$M)$——即如果 $0.98u(\$M)$ 大于 $u(X)$——那么只拿第二个盒子就比拿两个盒子的期望效用要更大。那么,基于效用与钱数是线性关系的预设,只要第一个盒子里的钱少于 98 万,则这个人就会选择只拿第二个盒子里的钱。所以,举个例子,在与纽科姆问题具有相似结构的一个选

---

① 一个例外是 J. H. Sobel, "Infallible Predictors", *Philosophical Review* 97 (1988);他在文章结尾考虑了"一个极限纽科姆难题",在其中所讨论的情形是第一个盒子中的钱数从一千美元增加到了(几乎)一百万。然而 Sobel 并没有继续考虑把一千美元减少到几乎为 0 美元的情形。Kenneth MacCrimmon and Stig Larsson, "Utility Theory: Axioms versus 'Paradoxes'",载于 *Expected Utility Hypothesis and the Allais Paradox*, ed. Maurice Allais and Ole Hagen (Dordrecht: Reidel, 1979), p. 393, 考虑了第二个盒子(尽管不是第一个盒子)中钱数发生变化的情形。

择问题中,其中的第一个盒子里面有 97 万 9 千美元,而第二个盒子一如既往地可能有 M 美元或者一分钱也没有,那么这个人也不拿两个盒子里的东西而只拿第二个盒子里的东西。无疑,钱的效用在此范围内并不是与其数量成线性的,但对于我们的目的而言,这一点并没有很大的影响——$M+X$ 的效用还是会成比例地低于它的钱数。尽管如此,其一般意义还是成立的:对于第一个盒子中的数额很大的钱,比如说 90 万,假设那个存在物能极准确地做出它的预测,那么第一个论证的倡导者会只拿第二个盒子中的东西。然而,在此种情形下,我们中极少有人会心安理得地遵循第一种论证,且承认另一种论证——我们拿两个盒子的东西在哪种情形下都会更好——不具有任何力量。

无论人们会支持何种论证来对初始纽科姆难题做出选择,通过变化第一个盒子里面的钱数,我们都能使得他们极为不安。当第一个盒子里的钱减少到 1 美元的时候,初始选择两个盒子的人不会再愿意遵守那个占优论证;当第一个盒子里的钱增加到 90 万的时候,初始只选择第二个盒子的人也不再情愿遵循那种期望效用的论证(具有不标出任何影响的条件概率)。这表明:对于初始的纽科姆例子,任何人对自己所遵循的观点都不抱有**完全的**信心。没有人愿意毫无保留地、全盘地应用似乎打动了自己的那种推理论证。 45

个人对于各种决策原则(和相伴随的论证)可能持有不同程度的信心。我们暂时可以把我们的讨论限制在最大化(条件)期望效用的两种原则内。这两种原则分别是由因果决策理论和证据决策理论所表述的。这些不同程度的信心可以由包含 0 和 1 及之间的(总和为 1)"信心度"所表示,也可以由总和低于 1 的信心度所表示(因为要保留这种可能性:即对于某一给定的情

形,两种原则可能都不正确),或者不在 0 和 1 之间的那种信心权重程度所表示。对于具体的某个人,我们用 Wc 代表他或她给予因果决策理论的期望效用原则的权重,用 We 代表他或她给予证据决策理论的期望效用原则的权重。用 $CEU(A)$代表行为 A 的因果期望效用,即指这一行为的效用能够依据(各版本中受支持的一种)因果决策理论而计算出来;用 $EEU(A)$代表行为 A 的证据期望效用,即指这一行为的效用能够依据证据决策理论而计算出来。每一种行为都将伴有一种决策价值(decision-value)DV,这是对其因果期望效用以及证据期望效用的一种加权价值(weighted value),其权重就是这个人对于受这两种期望效用之一所指导的信心度。

$$DV(A) = Wc \times CEU(A) + We \times EEU(A)$$

并且这个人会选择具有最大决策价值的一个行为。[①]

我认为,我们要更进一步,不仅仅是说不能确定这两个原则[$CEU(A)$和 $EEU(A)$]中**哪一个**(完全从其自身看)是正确的,而是说这两个原则都是合法的,且每一种都应当给予其应得的分量。那么,这种权重并不是不确定性的度量,而是每种原则之合法效力的度量。因此,我们就有了一种**规范**理论,指导人们去选择一个具有最大决策价值的行动。

如果一个决策价值最大化者赋予 Wc 和 We 的权重都不为零,那么这就会导致他在纽科姆难题中改变自己的选择:当第一个盒子中的钱数增到足够多时,他会从要一个盒子转变为要

---

① 如果对一项原则不具有完全的信任,就会使一个人去遵循各种原则的一种组合,那么如果他或她对这一组合不具有完全的信心,又会如何呢?如果这个人对于另外某个确定的原则抱有**些许**信心,那么,只要一个人的论证要取决于实际的信心程度,由此看来这一另外的原则也应当在权衡时被包括进来。

两个盒子；当第一个盒子中的钱数降到足够低时，他的选择又会从两个盒子转变为只要一个盒子。对于决策价值最大化者而言，这种转变也是可预测的。[因此，最大化 *DV* 的理论具有可检验的(testable)、定性(qualitative)行为结果，至少对于那些遵循这个规范理论的人而言是如此。] 46

有许多不同的数学结构都可以给予 *CEU* 和 *EEU* 一种作用，但 *DV* 公式是特别简单的一种，而现在就看更为复杂的东西则为时过早。当然，加权的 *DV* 结构就其自身而论并不能给予人们很多指导。权重应当有多大呢？人们必须在所有的决策情境下都使用同样的权重吗？还是针对不同种类的决策情境，权重是可以变化的呢？或者更系统地说，按照一种决策的情境落在某个维度 *D* 的哪个位置——越往左则使用这一种决策准则就越可行(因此该准则获得的权重就越大)，越往右就使用另一种准则越可行？我欢迎在贝叶斯结构内确定或限制那些"先验概率"(prior probabilities)的理论，也欢迎一种在通常的排序公理中确定或限定那些偏好的实质内容的理论，同样，我也欢迎一种确定或者限制权重的理论。在每种情形下，一般性结构仍然是有阐释力的。

两种因素(*CEU* 和 *EEU*)都要被赋予某种权重，意味着即使在行为没有对相关后果产生任何因果影响的决策之中，*EEU* 也会得到某种权重。比如说，职业生涯的选择表明(但不影响)将患上或者已经患上一种可怕疾病之不同概率的情形。在我的初始论文中，我认为给予这种考虑以丝毫权重都是荒谬可笑的。然而我还是知道，*DV* 公式中的证据成分对人类历史具有重大社会影响，正如加尔文主义和它关于得选标志(而不是原因)的观点，在资本主义发展过程中得到证实具有重大作用那样。(这可看成是一个行为的因果性结果："个人相信该行为"表明了但

不是引起了某些事情,并且他由于这个信念而感到很幸福。然而,若个人主张把这一点作为此行为的原因,那么他也必须注意这一点,即赞同把这种愉快的后果作为持有这一信念的理由。①)

理性的研究者一直试图建构一套唯一正确且完备的原则,从而毫无保留地应用于所有决策情形中。但是他们还未做到这一点——无论如何,我们对此没有完全的把握。在此种情形下,一个审慎而理性的人不是会两面下注吗?也就是说,我更想说的是,没有任何一个单独的原则是完全恰当的——这不仅仅是我们还没有为"何种原则是正确的"找到一个一槌定音的论证。47 我并不是说决策价值框架就会使理论者达成一致。即使他们在应包含何种原则在内这一点上达成了一致,但在各个原则应赋予多大权重上,仍是众说纷纭。正是这种权重分歧解释了对纽科姆难题的那些不同选择,但正是我们确实赋予**权重**(而非唯一忠诚于某一种原则)这一事实解释了决策会随第一个盒子里钱数的变化而改变。*DV* 结构表达了 *CEU* 和 *EEU* 各自都把握了(一类)合法理由,且我们不想完全忽视任何一类理由。②

让人多少感到怪异的是,决策理论的研究者们对他们的观

---

① 对于证据考虑的一种不同观点,即认为仅当它们匹配于人际间情境的合作推理时,它们才有吸引力,参见 Susan Hurley, "Newcomb's Problem, Prisoners' Dilemma, and Collective Action", *Synthese* 86 (1991): 173 - 196。

② "但是什么东西解释了 *CEU* 和 *EEU* 的倡导者之间的分歧呢?它是一种事实分歧还是价值分歧?"这个问题假定了两个倡导者共享 *EU* 公式,来追问他们的分歧是否在于概率成分或效用成分。然而,如果 *DV* 公式是正确的,那么就存在**其他**有分歧的东西,包括 W_c 和 W_e 的权重,以及这个公式的性质,还——预见了下面的段落——包括了其他的因素。追问"事实或价值?"——不允许再有其他选项——是假定了共同的东西**必定**是简单的 *EU* 框架,只在它**内部**才出现分歧。

点普遍表现出了这样一种信心。因为如果我们把有关正确的决策理论这一议题作为一个决策问题，亦即关于我们应遵循哪种决策原则的问题[①]——我们可以想象灵药已被发明出来，能使我们变成对于某一种原则持之以恒的履行者——那么互相竞争的诸种原则会给出何种答案，并不是显而易见的。尤其是，是否每一种原则都想要使自己成为更可取的选项，这一点也并非是显而易见的，这要取决于这个世界的情况。如果世界上有许多类似纽科姆问题的情形，且有极为重大的报偿，那么预计吃 *EEU* 药丸就会得到更好的因果结果，由此 *CEU* 原则将会建议人们去吃 *EEU* 药丸而不是吃 *CEU* 药丸。另一方面，如果世界上存在许多重大情境，它们都是具有我的疾病情形结构和类似情形[吉巴德(Gibbard)与哈伯(Harper)举的所罗门(Soloman)例子等][②]的结构，那么遵循 *EEU* 原则的人们不做任何“提醒(tickles)”会经常错失重要的利益(因为它们所预告的不幸)。既然这点能够预测得到，那么 *EEU* 原则本身就会推荐人们去吃 *CEU* 药丸，这个行动要比吃 *EEU* 药丸具有更高的 *EEU* 效用。(在这种情况下，*CEU* 原则也会建议人们吃 *CEU* 药丸的。是否会有这样的例子，那里吃 *CEU* 药丸具有更高的 *EEU* 效用——由此 *EEU* 原则建议这样做——而此时的 *CEU* 效用却是更低的，因而 *CEU* 原则不建议人们吃 *CEU* 药丸？那样的话，此问题将困难重重。)正如不存在任何一种特定的归纳策略，即没有任何一种卡尔纳普式 c-函数，它在任何世界里都是最好的或最有效的，所以也就没有一种最好的

---

① David Gauthier 考虑了个人应该选择具有何种选择倾向的问题，见 *Morals by Agreement* (Oxford: Oxford Univ. Press, 1985), ch. 6, secs. 2-3。

② Nozick, “Newcomb's Problem”, p. 125; Gibbard and Harper, “Counterfactuals”, 也参见 p. 000n2。

合理决策原则。① 而且正如我们想使我们的归纳程序允许学习,亦即包含由对世界的经验而确定的那种参数,我们也想我们的合理决策原则能包含这样的参数,它们契合于在我们要做决策的世界里所发现的各种特征。(在每种情形下,在设定参数以48符合真实世界这一过程中,进化可能起了重大的作用,但这并不意味着,我们应当指望我们具体的归纳策略或决策原则适用于每一种可以想象的科幻情境,或者我们应当把它们作为先验有效的来对待。)决策价值的框架,其中所纳入的权重是可以随时间而改变的,是实现与现实世界相契合的一种方式。

我们确定的决策价值是以 CEU 和 EEU 成分为基础的,但是还可能存在其他可行的、可选的决策原则或决策因素,它们也许会以具有相关权重的条款而加上去。特别是,我们可以把一个行为的象征效用(即它的 SU,它纳入了该行动所象征的各种结果与行动的效用)以权重 Ws 添加到公式中来。(最好是不要纳入伴随其他效用的象征效用,因为它很可能不符合一种期望效用公式,也因为我们想要象征效用保持分立的轨道,既然我们认为在不同类型的决策情境中给予这一因素以不同的权重是合适的。)那么,A 的决策价值($DV(A)$)公式就产生了:

$$DV(A) = Wc \times CEU(A) + We \times EEU(A) + Ws \times SU(A)$$

研究这个决策价值结构的形式特征是颇有教益的。如果像过去有关不确定性决策的文献中所探讨的那些准则那样,

① Rudolf Carnap 强调[*The Logical Foundations of Probability* (Chicago: Univ. of Chicago Press, 1950)],断言"基于 e 确证 h 的程度是 n"的各种句子是分析性的,如果是真的话。然而,他也认为在选择何种特定的确证函数(c^*, c^+ 或任何归纳方法连续统的函数),因此何种函数将规定这种分析性关系,是个实践选择的问题,它将取决于这个宇宙的一般事实。

有时这一加权组合原则不能展示出某些可欲的特征，这是不足为奇的。①

象征效用并不是一种不同类型的效用，它与事物的标准效用之间的关系不像比喻意义与字面意义之间的关系。确切地说，象征效用与常见类型效用之间具有一种不同类型的（象征的）关联。它与那些常见的因果和证据关联相并列。正如 A 的 CEU 是由 A 与具有标准效用的结果之间的那种因果概率关联所决定的一样，行动 A 的象征效用是由 A 与各结果（还可能是其他行为）之间的象征关联所决定的，这些结果或行为自身是具有标准类型的效用的。②

49

我们应当确保这些类型的关联——因果的、证据的和象征的——都是互斥的吗？早前的 DV 公式把它规定为一个由 CEU 和 EEU 的加权和。然而一个行为的 EEU 包括了它的因果性成分，因为既定行为之结果的条件概率[即 EEU 理论家所使用的 $prob(O/A)$]在存在因果影响的情况下纳入了它。那么，在我们的加权和公式中，我们不应该把 EEU 解释为由那些

---

① 参见 John Milnor, "Games Against Nature", 载于 *Decision Processes*, ed. R. M. Thrall, C. H. Coombs, and R. L. Davis (New York: John Wiley, 1954), pp. 49-59, and R. D. Luce and Howard Raiffa, *Games and Decisions* (New York: John Wiley, 1957), pp. 275-298。我早前说过，象征意义不一定成比例地进入概率语境。然而 DV 公式把象征效用作为一种加权成分而包括进来。我们也许想知道象征效用是否要进入加权 DV 语境中去。然而，向概率情境的转变是向不同情境的转变，然而，转变到 DV 公式并不会转变选择情境。

② 因此，对于在 59 页注①脚注中所讨论的度量效用的标准情境而言，我们有必要施加一个进一步的条件。该情境必须是这样的，那里行为与其效用结果之间不具有任何相关的证据关联或象征关联。在行动与有价值结果之间具有证据和象征关联的地方，因果语境中的效用就要得到校正，且要遵循一种期望效用原则，并且要使用由此发现（校正后）的效用。然而，后面这些结果的价值是在完全因果语境下度量的。

不是(仅仅派生于)因果概率(部分)的那个概率(部分)所表现出来的期望效用吗？而且类似地，行为的象征效用 *SU* 不应当是那些并非(仅仅)派生于且在其中表达的那些因果和证据关联的那种象征效用吗？[1]

我们还应当在 *DV* 结构里进一步地纳入其他成分吗？有个建议是明确地纳入一种成分，它与行为符合个人的自我形象和自我表达的那种方法相关。事实上，这个领域之中的很多东西已经涵盖在我们的三个成分里了。尽管做出某一类人常会做出的那类行为可能不会**致使**(cause)那个人成为那类人，但这可以象征他是那个样子的，即作为他是那类人的某种证据，且具有这样一种因果后果，即让他能够更容易维持自己是那类人的一种**形象**。最后这一点是一个行为的真实的因果后果，可以具有很重大的效用。因此这种后果，即某一具体行为如何影响了该人的自我想象——能够在他的行为解释理论中起到一个重要的、明确的作用，即便行为人不能轻易地将这种后果明确地纳入考虑(“我要去做 *A* 是为了我自己更易于维持‘我是一个 *K* 类人’的那种自我形象”)，而不致因此会降低这一效果本身。[2] 我们应当认为，“表达性(expressveness)”不是这些狭隘地设想且独

50

① 我所知道的既处理因果关联，也处理象征关联，且试图把它们分离的一个心理学研究是 G. A. Quattrone and Amos Tversky, “Causal versus Diagnostic Contingencies: On Self-Deception and the Voter's Illusion”, *Journal of Personality and Social Psychology* 46 (1984): 237 - 248。

② 并不是每一种行动模式都包含一种与后果的联系，它可以置于根据因果的、证据的和象征的联系的公式内。考虑无动机的行动，佛教、道教和印度教中的文献中谈到这种模式的各种变体。这里个人行动不是为了变成某个样子，或成为某个样子，或是为了产生各种结果，或者是为了具有证据，或者是为了象征任何东西。也许他这样做只是为了把他本人(正确地)与最深刻的实在联合(align)起来，或者让它对他起作用而被最深刻的实在联合。这种行动模式需要进一步的分析，但是它看起来并不包含一种与后果相关联的模式。

立的其他范畴所能穷尽的,因为正如我先前说过的:象征的与表达的这类范畴是交织在一起的。

如果我们把语言学范畴与我们已有的类别排列如下:因果的/证据的/语言的/象征的,那么这提出了两个问题。象征范畴(象征范畴确实更适于效用的归回,但这并不一定每次都发生)如何在本质上不同于单纯的语言学范畴呢?并且语言学范畴如何能够产生于因果范畴和证据范畴?当因果关联和证据关联(它们来自因果和统计规则性的一个分支结构)是常识的时候,为了使另一个人相信 p,某人也许会有意地制造出 p 的一种证据标志。这是超越格莱斯式(Gricean)自然意义(它只是一个证据标志)的关键一步,从而在另一个人那里有意地使用它来产生一个信念,也就是说,在部分程度上走向了格莱斯的非自然意义,这里意向性(intention)是有意得到承认的。[①] 这样一种证据标志也许是用来诱发人们对于一个真 p 的信念,另外那个人在当时是不可能独立地观察到的。但也许同样可能的是,当 p 为假时,基于一个安排好的证据,它首先是用来欺骗另一个人相信 p 的。代表其他事物的第一个陈述也许是一个谎言,一个伪造的自然标志。如果语言界定了人性,即表达了人类的理性能力且把人区别于动物,那么对于"我们生来就有原罪"这一学说而言,这将带来一个颇有意味的转变。

囚 徒 困 境

囚徒困境是个得到了很多讨论的情境。在此情境下,如果每一方都选择对自己来说是(强)占优的行为(那看上去是要做

① H. P. Grice, "Meaning", *Philosophical Review* 67 (1957): 377 - 388.

的合理事情),其结果相对于选择更合作的行为而言,每一方都会变得更差。他们的个别理性组合在一起,反而使他们失去了一种本可获得的更好情境,因此这是帕累托次优的(Pareto-suboptimal)。

这个一般情境是依其一个例子而命名的:一位司法官向两位等待审判的囚徒提供了下列选项。(这个情境在两个囚犯之间是对称的;他们不能沟通来协调他们的行动以回应司法官的出价,或者,如果他们能的话,他们也没有任何手段来强制执行他们达成的协议。)如果一个囚犯坦白而另一个不坦白,那么第一个人就不用服刑,而第二个人将被判入狱 12 年;如果两个人都认罪了,那么每个人获罪服刑 10 年;如果两个人都不坦白,那么每个人将被判罚两年。(表 2 表达了囚徒所面临的情境,矩阵中的条目分别代表了第一个和第二个囚犯将要服刑的年数。)

表 2

		囚徒Ⅱ	
		不坦白	坦　白
囚徒Ⅰ	不坦白	2,2	12,0
	坦　白	0,12	10,10

每一个囚犯会做如下推理:“如果另一个人坦白而我不坦白的话,我会入狱 12 年;然而如果我也坦白,我会入狱 10 年。如 51 果那个人不坦白我也不坦白,我会入狱 2 年;而如果我坦白,我根本就不用进监狱了。无论在何种情况下,无论另一个人怎样选择,我坦白都比不坦白要好。因此,我要坦白。”每个囚犯都以同样的方式推理:都坦白且都获刑 10 年。但如果他们都不坦白,每个人都只获刑 2 年。个别理性的组合造成了一种共同的

混乱。并且这个情境在如下意义上是稳定的：两个人都没有动机去做其他(更合作的)行为，既然另一方打算坦白的话。他们的坦白行为处在均衡状态中。

囚徒困境情境例示了一个更一般的结构(见表 3)，那里每方都可以选择两个行为之一——把占优行为称作 D，把合作行为称为 C——并且他们对组合行动的可能结果 a、b、c 和 d 之间具有下列偏好。个人 I 的偏好降序是 c、a、d、b，而个人 II 的偏好降序是 b、a、d、c。既然个人 I 偏好 c 于 a，d 于 b，所以行为 D 占优于行为 C，因而他选择去做 D。因为个人 II 偏好 b 于 a，d 于 c，所以行为 D' 占优于行为 C'，因而她选择去做 D'。D 和 D' 一起产生了结果 d，但是两个人偏好结果 a(从对 C 和 C' 的选择中会产生这样的结果)于 d。因此，有关这个 2×2 矩阵的结构和每个人的偏好序列结构的事实是简单的，但足够标出一种囚徒困境的情境。

表 3

		II	
		C'	D'
I	C	a	b
	D	c	d

有些人主张，此情境下的一个理性者，他知道另一个人也是理性的，且像他一样理解这个情境，因此他认识到，任何能够说服自己的推理也同样会说服另一个人。这样一来，若他的结论是占优行为是最好的，那么另一个人也会这样认为；若他的结论是合作行为是最好的，那么另一个人也会这样认为。那么，在这种情境下，得出“合作行为是最好的”是更好的，因此意识到这一切之后，他会做出合作行为。这类论证是褒贬

参半的。

不管囚徒困境与纽科姆问题是否(像有些人所争辩的)在所有本质特征上面都一模一样,两者都是可相提并论的。两者都涉及会导向不同行为的两种论证:一种论证立足于占优原则,它是用一种适宜于因果决策理论的方式解释的。另一种论点则是立足于以契合于证据决策理论的方式来考虑每个行动所指明的东西(因此是应该据之打赌的结果)。在囚徒困境中,尽管你的行为并不会因果地影响到另一个人的行为,但你预计他会做你所做的那种行动,这个论证是最大化证据期望效用的原则所允许的,那里条件概率不需要表达任何因果性的影响。因果决策理论会建议做占优行为;当你认为另一方与你在相关方面类似时,证据决策理论会建议做合作行为。这毋需你确定你们双方都会做出一样的行为;给定你的行动,只需要另一方行为的条件概率能够发生充分的变化,这就足够了。(也请注意,如果你认为另一方将会做出不同于你的行为,或者是因为她的行为独立于你自己的行为,且你对她合作的概率做出足够悲观估计的话,证据决策理论也许会导致占优行为。)正如在纽科姆难题的例子中那样,我们对两种观点都没有完全的信任,我们可能想要对每个立场都给予某种合法的权重。

在纽科姆难题情形中,随着第一个盒子中钱数变化而出现的决策转换,这种合法权重的多重认可(或者换种说法,是缺乏完全的信心)表明了其自身。(然而,由两种相冲突的决策原则所判断的这种问题结构仍然保持不变,这些原则在钱数改变中会保持同样的决策。)在囚徒困境情形中,问题在于当理性行为者都知道自己面临的是理性行为者时应该做什么。两种不同论证的倡导者发现,表 3 中的抽象结构足以对

他们所偏爱的论证给出有说服力的理解。占优论证所需要的一切是：个人 I 偏好 c 于 a，d 于 b，并且个人 II 偏好 b 于 a，d 于 c。对理性行为者而言，证据期望效用论证看来需要的一切是：每一方都知道彼此是理性行为者，且彼此都偏好 a 于 d。然而，如果人们确实对这些观点缺乏完全的信心，那么我们就会发现，随着表 3 的抽象结构中的各种报偿数量[①]的变化（行为者的偏好顺序仍然保持不变的情况下），人们做出的决策也会随之变化。

53

假设效用是以等距尺度来衡量的，单位是任意的，零点也是任意的，但是遵从标准的冯・诺伊曼-摩根斯坦公理的某种变体，且可做正的线性变换。[②] 在表 4 所表述的情形中，矩阵条目是这样的效用数目，据此我们会认为合作行动是合理的选择。一般而言，当合作选项的回报远远高于占优选项的回报，且不匹配行为的回报对两者提供的是微不足道的利益或者损失时，那么我们会强烈地认为合作行为是合理的，并且认为占优论证是没有什么力度的。换言之，在表 5 中，合作解只是比占优解略好一点，并且不匹配行为的报偿之极值差距甚大。当我们与另一方没有特殊的纽带，且对另一方的行为概率并没有详尽的了解时，那么我们会认为在表 5 情境下做占优行为是合理的，而不冒另一方做占优行为给我们带来的风险，而这是他有很强的动机去这样做的。（如果我经历了这一推理，且认为他很可能会像我

---

① 更准确地说——因为效用以等距尺度来衡量——是不同数量之间的差别比率。下面的讨论为了明晰性而忽略了这种复杂性，它是可以做出恰当的重述的。

② 参见 John Von Neumann and Oscar Morgenstern, *Theory of Games and Economic Behavior*, 2d. ed. (Princeton: Princeton Univ. Press, 1947)，附录。对于有关 Von Neumann-Morgenstern 和类似的条件组的哲学问题的考察，参见 Robert Nozick, *The Normative Theory of Individual Choice* (PhD. Diss, Princeton University, 1963; rpt. New York: Garland Press, 1990)。

一样，那么我在此情形下就会确定做占优行为，并安心地认识到他也会这样做。）

表 4

		II	
		C′	*D′*
I	*C*	1 000，1 000	0，1 001
	D	1 001，0	1，1

表 5

		II	
		C′	*D′*
I	*C*	3，3	−200，500
	D	500，−200	2，2

个人做出的决策会随矩阵中具体的效用条目的数量（的差距比例）而有所变化，而这种变化是与早前人们赋予 *CEU* 与 *EEU* 一定权重的那种最大化决策原则相符合的。至于当效用变化时，人们的决策会确切在哪个点上转变，这既要取决于她对每个原则有多大的信心（也就是说，她暗暗地赋予了这些原则以多大的权重），也要取决于她认为另一方与她做出同样行为的概率。然而，要注意到，即使后者的概率为 1，即使行为人赋予 *EEU* 的权重比 *CEU* 的更大，她仍然不是必然会选择合作行为。如果效用筹码足够大，并且符合表 5 中的情境，再结合行为人赋予 *CEU* 原则的权重，或者结合占优原则本身（以某种因果变体），或者结合要重视一定安全水平的其他原则，就会推荐采取占优行为。即使你绝对相信另一个人会做出与你一样的行为，这也不足以确保你会做合作行为——如果 *EEU* 原则不具有绝

54

对信心或绝对权重的话。①（我到现在一直假定，它是一个人在所有决策情境下都会应用的 DV 原则的一个特定的版本，有着固定的特定权重。然而，情形也有可能是，对于决策的一套构成原则而言，随着行为者所面临的决策情境的类型变化，她也可能赋予那些原则以不同的权重。尽管如此，只要得到正权重的具体原则不止一个，总有这样或那样的 DV 结构是契合于这类情境的。）

在前一节中，除了 EEU 和 CEU 之外，我们在 DV 结构中还纳入了做一个行为的象征效用，即它的 SU。也许有人认为，如果行为确实具有象征效用，那么这会在那个行为的效用矩阵条目中将自身完全展示出来（例如，也许每一条目都会增加代表该行为之象征效用的某一确定的数量值），因此不一定存在任何独 55

① “当另一个人与你做出同样行为的概率是 1 时，你就应当在囚徒困境的情形下选择合作的行为，而不管矩阵之中的效用数量的差别有多大，一种正确的理论难道不会这样坚持吗？那么这种异议不是对于 DV 结构的一种反驳吗？”但是，我们怀疑这个人对他所做出的这个概率为“1”的预期（在一个更高的层次上）是否具有完全的信心，且缺乏完全的信心是否会影响到他在高风险境况之下的行为？参见 Daniel Ellsberg, “Risk, Ambiguity, and Savage Axioms,” *Quarterly Journal of Economics* 75 (1961): 643 - 669。

还要注意到这个论证进行得过快，从(1) 共同的理性；到(2) 他们将做出同样的行为；到(3) 划掉矩阵中代表着不一致行为的右上方与左下方的这两个矩阵条目，它们代表着有分歧的行动；再到(4) 主张，既然是在余下的两个盒子中进行选择，双方都应当选择他们偏爱的那一个——各自所偏爱的——即双方都应当做出合作的行为。假设对理性的共识允许我们假定，我们会以同样的方式推理，并且最终会做出同样的事。但也许那也是源自每个人对矩阵中四个盒子进行推理，并且各自根据由整个矩阵（包括四个盒子）所代表的联合策略得出，我（以及他或她）应当选择不合作的行为，于是两个人最终得到的是右下方的代表不合作的盒子之中——由此也满足了他们做出相同行为这一条件。事先知道我们将会做出同样的事，此时就意味着我们知道，我们的结果不会处在右上方或左下方盒子里。但这并不意味着我们因此可以先删除它们，然后再对余下情境进行推理。因为或许我们最终会得出做相同行为的推理取决于不先行删除那些不一致的角落。

立的 *SU* 要素。然而,一个行为的象征价值并不是由**该**行为所唯一决定的。该行为的意义可能取决于:你还有什么其他选项,它们具有何种报偿;其他人还有什么其他选项。该行为所象征的东西是:在**那个**具体情境中,优先于**那些**特定选项的情况下做它而象征的东西。如果一个行为象征着“成为一个合作者”,它有这样的意义不仅是因为它具有它确实具有的两种可能报偿,同时也是因为它在那个二人矩阵中具有特殊的地位——即做一个**严格劣势**(dominated)行为(当与另一个人的**严格劣势**行为联合时)比占优行为的组合对每个人都产生更高的报偿。因此,它的 *SU* 并不是通过孤立地处理该行为而把握的那些特征的一种函数,即不仅仅是从状态到结果的一种映射。[①] 行为的象征价值可能取决于整体决策或博弈矩阵。它的象征价值无法通过在矩阵**内**的效用结果上做加减而恰当地表达出来。许多论者假定**任何东西**都能够形式地嵌入到后果之中,[②]比如,做一种行为有何**感觉**,你已经做了该行为这个事实,或者它属于特定的义务论原则这个事实。但是如果去做一个行为 *A* 的**理由**将影响行为的效用,那么试图把 *A* 的这种效用嵌入其**结果**,这就会改变这个行为,进而改变做这个行为的理由;但是**那个**被改变的行为之效用将取决于做**它**的理由,而试图将此嵌入其结果又会改变去做这个现在已被改变过两次的行为的理由,以此类推。不仅如此,若行为是出于某些理由而做的,那么其**结果**的效

---

① 这就是 L. J. Savage 在他的决策理论的形式论中如何处理行动的;参考他的 *The Foundations of Statistics* (New York: John Wiley, 1954)。然而,即使不管有关行动的象征价值等问题,它也不能被如此还原。参见我的 *The Normative Theory of Individual Choice*, pp. 184 - 193。

② 例如,参见 Peter Hammond, “Consequentialist Foundations for Expected Utility,” *Theory and Decision* 25 (1988): 25 - 78。

用就能改变。[①] 因此,我们想让后果效用表达的东西是,给定该行为是出于某些理由而做的,该结果所具有的条件效用。[②] 这就给后果主义在处理动态一致性问题上造成了麻烦;因为情形也许是,达到了决策树中的一个具体分支(subtree)的这一事实给了你改变一种未来结果之效用的信息。如果我们试图坚持用"决策树内的效用总是完全确定的条件效用"来应对这一点的话,那么我们在决策树中任何两个不同的位置上就不可能具有**同样的**结果——这不利于陈述一般性的规范原则来管辖这种决策树。(对于行为的**每一个**事实,也许都有一种描述方式使你能够把这个事实列为该行为的一个结果,但这不能得出存在这样的描述,以致对于该行为的**一切**事实而言,该描述都能把它们纳入该行为的后果中去。量词的排序是重要的。) 56

这些考虑表明在囚徒困境的情境下,行为应被设想为其自身具有一个效用,而不是简单地作为矩阵行**之内**包含的一个固

---

① 由于纽科姆难题,我们探讨了这样的情形,其中结果的**概率**随着做这个行动的理由而改变,由此引出了关于"可核准性(ratifiability)"的文献。

② 甚至就是这个后果的条件效用,给定行为是出于某些理由而做的,**且导致了这个结果**。在拍卖的经济学文献中,人们指出,当一个人发现他的出价获胜时,且当这标明其他有见识的出价者有信息或得到的结论导致他们没他这么看重这个结果时,这个人自己对该后果价值的估计也会发生改变。可核准性方面的文献注意到,"我决定去做 A"这一事实能够影响到对 A 的结果 C 的概率估计,因为 $prob$(C/我决定去做 A)并不等同于 $prob(C)$,同时拍卖方面的文献指出,"我做 A 成功地引起了 C"能够影响 C 的效用,这可能是通过改变能影响 C 之效用的其他信息的概率而达到的。因此,一个构造完全的决策理论不仅必须要运用条件效用——见拙著 *The Normative Theory of Individual Choice* (1963; rpt. New York: Garland Press, 1990), pp. 144 - 158——而且它所利用的条件效用必须不简单地是 u(结果 O/完成的行为 A),相反是 u(结果 O/完成的行为 A,出于原因 R,且出于 R 而做的 A 产生了 O)。

定的效用增加。[①] 但我希望做出更强的主张：这种效用是一种象征效用。我并非简单地意指应用于一个行为而非一个结果的那种普通类型的效用。这一效用涉及一种不同类型的关联。在某些囚徒困境情形中，做严格劣势行为——通常这被叫做“合作行为”——对该人可能具有象征价值。这可能代表了他在与他人的互动中是合作者，亦即在互利的联合风险事业中，他是一个自愿且不挑剔的参与者。那么，这种情境下的合作就可能与那些并没嵌入在囚徒困境情境下的其他合作行为归为一类。因此，在这一特殊的囚徒困境情境下不合作，就可能进而威胁到在其他情形下也不合作——界线未必如此分明，他在其他情境下的合作动机也可能在部分程度上是象征的。因为他赋予成为一个合作者以很大的效用，那么在某一特殊的囚徒困境情境下，他会做出象征这一点的严格劣势行动。[②]

这并不意味着行为人将只顾及该行为的 *SU*。他也会考虑行为的具体效用条目和这些条目如何受到 *EEU* 和 *CEU* 原则的评价。行为对他的决策价值将取决于所有这三种因素——行为的 *SU*、*EEU* 和 *CEU*——和他赋予这些因素各自的权重。因此，他赋予做一个合作者以某种(正的)象征效用，单单这一事实并不足以保证他会在所有囚徒困境的情境里都做合作行为。

57

我并不主张，与囚徒困境的情境相关的唯一可能的象征意义是“做一个合作者”。有些人也许认为，在这种情形下做占优

---

① 也值一提的是，当行动的序列在策略上是相关的时候，博弈论者不能简单地聚焦于一个博弈和其报偿的矩阵表达，而要考虑博弈树。

② 行为人在囚徒困境情形下做占优行为，其恒定结果就是他将认为自己是个不合作者，然后通过博弈矩阵之中加一个负值而表达出来，亦即在表示不合作行为的每行条目中都加上一个负效用，我们可以这样说而将此嵌入到标准决策理论当中去吗？要注意当这个效用表达在整个矩阵结构之内时，效用的这个成分就会成为他对该行为态度的一个函数。

行为象征“做不为情感所左右的理性人”。如果认为这一点很重要的话，那么他就会(在他的 *DV* 原则之内)赋予做占优行为以很高的象征效用，这是在他给予 *CEU* 或占优原则本身的权重之外的。在纽科姆难题中，有些论者支持“拿两个盒子里的东西是最合理的选择”这种观点。若他们和类似他们的人面对比最大化 *EEU* 者在这个问题上做得要差这个事实，他们克服这种不安的说法，其“要旨”在于：“若某个人非常善于预测，且给予所预测到的那种不合理性以丰富的报偿，那么那种不合理性就将是有丰厚报偿的。”①我认为这些人(根据他们对“做理性人”包含何种确切原则的最优估计)赋予了“做理性人”以非常高的效用——这是一种象征效用吗？[一个微妙的问题是区分这样的两个人：一个人只赋予一个特定的原则(比如说 *CEU* 原则)以权重；另一个则是赋予 *CEU* 原则一定的权重，同时也赋予 *EEU* 较小的权重，而且赋予了遵循她对合理性内容的最好的具体评估以很大的象征效用——很大的权重。]若遵循特定的决策原则本身具有象征效用，或者从事具体类型的决策过程或程序也有象征效用，那么人们就会预计出现新的复杂性。

对于象征效用所说的这一切表明：我们对于囚徒困境的回应，在部分程度上是受到我们想做哪类人和我们想与他人具有何类关系所支配的。我们在某一具体的囚徒困境情形中所做的将包含所有这一切，并且在不同的程度上援引它，其具体程度则既取决于矩阵中精确的效用条目(其间的差别比例)，也取决于引起该矩阵的特定事实环境，行为在这种环境下可以逐渐获得自己的象征意义，而不仅仅是因为该矩阵的结构。

---

① Gibbard and Harper, "Counterfactuals and Two Kinds of Expected Utility", p. 151.

当然，我们早已知道所有这一切，至少是人们为什么在囚徒困境的情境下会做出不同回应的一种心理学意义。然而，*DV* 原则为“做何种人”的一般性观点留下了空间，因为这与具体的选择相关且要把它们归类，不仅仅是作为对（有些）人为什么会偏离理性的一种可能的**心理学**解释，而且是作为在他们的**合理**决策程序之内的一个合法成分，即象征效用。

在一篇讨论重复囚徒困境的重要文章中，克雷普斯（D. P. Kreps）、米尔格洛姆（P. Milgrom）、罗伯茨（J. Roberts）和威尔逊（R. Wilson）表明，如果你认为我做合作行为的概率很小，或你认为我相信你做合作行为的概率很小（或者你认为我相信你相信我会做出合作行为的概率很小），那么为了鼓励合作的行为或一致的信念，这就足以使得你率先做出合作的行为是合理的。[①] 如果你相信我会做出合作行为（或与你的行为亦步亦趋），而且如果你相信只有当你以某种方式行为时，我才会继续这样做，那么为了鼓励我做出合作的行为，你就有理由如我所想的那样去行为。[②] 如果这种情形是相互的，那么我们俩（在特定的环境下）都会做合作行为。现在若“大家都遵循 *DV* 结构”是种共识，那么 *DV* 结构确实（允诺）给每个人以某种概率相信另一个人会相信第一个人会做合作行为，由此使得每个人，也就是双方都有某种概率会做出合作行为。［要注意，这一观点和这段余下的部分并不取决于完整的 *DV* 结构，因为 *DV* 结构中还包

58

① David P. Kreps, P. Milgrom, J. Roberts, and R. Wilson, "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma", *Journal of Economic Theory* 27 (1982): 245-252.

② 正如一个作者总结道，博弈者也许会“打破一个均衡行动以打动另一个博弈者打破信念和策略的均衡”。Eric Rasmussen, *Games and Information* (Oxford: Basil Blackwell, 1989), p. 111.

括象征效用的权重。最初提出的狭窄 *DV* 结构(只考虑了 *CEU* 和 *EEU*)就够用了。]而且这一结果,不是作为对完全理性的背离,也不是对于其他人背离理性的(或是对其他人相信你会做出背离合理性之事的)一种合理调整,而是作为"所有各方都是彻底理性的"这一共识的一部分。因为如果最大化决策价值的原则是一种合理原则,即在规范上是可欲的,那么如果(正如它看上去的那样)*DV*-最大化的共识使得每一方以某种概率来做合作的行为,那么克雷普斯、米尔格洛姆、罗伯茨和威尔逊的论点甚至在完全理性的共识下也是适用的。①

如果因果的、证据的和象征的效用相互作用会导致得到比做合作行为更明确的结论,这就太棒了。在何种条件下,一方(或双方)参与者在 *DV* 结构中确定何种权重,个人才会在囚徒困境的情境下选择合作的行为,或者在重复式囚徒困境中采取一种针锋相对的策略呢?②

这里,我们只能采取临时性的初始步骤,列出恰当的假设来推出结论。除了要求参与的双方都遵循 *DV* 原则之外,我们还可以补充一个极弱的假设,即每一方应当料想另一方会像自己

---

① 顺带做一个附注。在我 1963 年的博士论文中,我看到了博弈情境中知识层次必然是无限扩展的,每个博弈者都知道博弈情境的结构,每个人都知道另一个人也知道,每个人都知道其他人知道他知道,依此类推(*The Normative Theory of Individual Choice*, p. 274)。但是我认为这只不过是吹毛求疵而已。我看不到理性的这一共同知识的深远利益和影响何在。参见 Robert Aumann, "Correlated Equilibrium as an Expression of Bayesian Rationality", *Econometrica* 55 (1987): 1-18 和 Drew Fudenberg and Jean Tirole, *Game Theory* (Cambridge, Mass.: M. I. T. Press, 1991), pp. 541-572。

② 对于针锋相对的策略的讨论,参见 Robert Axelrod, "The Emergence of Cooperation among Egoists", 重刊于 *Paradoxes of Rationality and Cooperation*, ed. Campbell and Sowden 和他的 *The Evolution of Cooperation* (New York: Basic Books, 1984)。

一样行为，并将遵循此假设填入 *EEU* 成分之中。这一弱预测 59 原则提出，另一方会做行为 C'的证据性条件概率（以你做行为 C 为条件）要比做 C'的无条件证据概率要大；这对于她以你做行为 D 为条件而做 D'的行为也是一样的。稍微强一点的一个原则，但还达不到对方会做与你的行为完全一样的行为的对称假设，它认为这些证据条件概率对于第一方而言是大于 1/2 的。另一原则规定行为者赋予在囚徒困境的情境中做合作行为以某种象征效用（和某些象征权重）。不仅如此，我们还可以假定做占优行为 D 不仅没有合作的正象征效用，本身还带有一种负的象征效用。① 给定另一方做行为 B，用 $S(A/B)$来表示行为 A 的象征效用。如果个人 I 赋予"做合作行为"以正的象征效用，那么 $S(C/C')$就大于或等于 $S(C/D')$，而这两者都会大于（负值的）$S(D/D')$，而它本身又会大于（更大负值的）$S(D/C')$。当囚徒困境结构在相同的双方之间重复进行多次时，那么互利的合作行为就更有可能影响到在当下这次博弈之中的效用，也包括最初第一次的博弈。不仅如此，一个行为的象征效用在各次博弈之间会发生变化，这取决于另一方过去的作为。我们会看到，当另一方越是做出占优行为，那么做合作行为的象征效用就越低，也许与另一方做占优行为与做严格劣势行为的比率成比例地降低。她拒绝合作的次数越多，选择与她合作对于"成为一个合作的人"的象征就越少。另一方面，另一方合作的次数越多，你做出合作行为所具有的象征效用就越大。一个可比较的条件也适用于做占优行为所具有的负效用。这一负效用的绝对

---

① 我在此直观地说，既然是讨论等距尺度的测量，具有一个任意的零点，那么所测量的数值为负就没有特别的意义。单位和零点是任意的，这为效用的人际比较带来了麻烦。我提出了一些建议，参见我的 "Interpersonal Utility Theory", *Social Choice and Welfare* 2 (1985): 161 - 179。

值随另一方做出占优行为时减少，随做合作行为时提高。我希望这些条件与其他可行的假设将得出更明确的结论。

更精细的区分：结果与目标

我们已经讨论了行为与结果之间三种不同的关联模式：因果的、证据的和象征的。我还指出，决策理论需要使用和明确承认所有这三种模式。决策理论在这些范畴之内还需要更精细的区分吗？例如，有些伦理学者主张不同类型的因果关联在各选择情境下携有不同的权重，即使引致的概率是完全一样的。他们主张，在引起某事件、允许它发生和不去阻止该事发生之间存在着重大区别。（而且，我们也许想到进一步的因果关联种类，诸如推动或帮助某事件的发生。）还有些论者构造了一种“双重效果”学说。这一学说认为在以下两者间存在一种道德上的差别（有时足以成为行为是否被允许的差别）：引起某事发生是源于将它作为一个目的或作为一个目的的手段而有意导致的，还有一个是明知会引发但却是作为追求其他某个目标的副效果而产生的。这些显然都是有争议的问题，①然而令人惊异的是，因

60

① 参见 Philippa Foot, “The Problem of Abortion and the Doctrine of the Double Effect”, 载于她的 *Virtues and Vices* (Berkeley: Univ. of California Press, 1978), pp. 19 - 32; Judith Thompson, “Killing, Letting Die, and the Trolley Problem”, and “The Trolley Problem”, 载于她的 *Rights, Restitution and Risk* (Cambridge, Mass.: Harvard Univ. Press, 1986), pp. 78 - 116; Warren Quinn, “Actions, Intentions, and Consequences: The Doctrine of Double-Effect”, Philosophy and Public Affairs 18 (1989): 334 - 351; Warren Quinn, “Actions, Intentions, and Consequences: The Doctrine of Doing and Allowing”, *Philosophical Review* 98 (1989): 287 - 312; Frances Kamm, “Harming Some to Save Others”, *Philosophical Studies* 57 (1989): 227 - 260。

果决策理论对这些按理来说很重要的区分至今还未曾给予任何关注;它反过来只是以一个不加区分的"因果影响"概念而进行的。无论是在第一人称选择理论(first-person theory of choice)中,还是在对建议者的教导中,规范决策理论应该为这种区分发挥作用而留有空间吗?这些区分能自然进入之处也许就是条件效用概念。早前在谈到拍卖理论的时候,我们注意到决策理论应当使用的是 u(结果 O/A 已被完成并且 A 引起了或成功地导致了 O)。这个条件的最后一部分中的行为与结果之间因果连接的确切类型会影响到引致的结果 O 的效用,也就是说,会对 O 产生不同的条件效用,因此在应用这些条件效用的原则之内有时会产生不同的决策。或者这些区分的要旨完全只是象征的,因此在我们已有的理论中纳入象征效用就已经给予了这些区分以足够的地位?①

我建议不要把这些区分看作是二分法,而要看成沿着一个(并不必然是连续的)向度而排列的项目。实际上,我们在此具有的不是一个向度,而是两个。第一个向度涉及行为的因果作用与效果、后果或引致的事态之间关系的重要性。这里一个行为与一个事态可以处于(至少)7 种关系中。随重要性的降低,一个行为可以:(1) 引起(Cause)这种事态的发生;(2) 帮助或推动它的发生;(3) 移除某个阻止它发生的障碍;(4) 容许或允

61

① 或者反过来,这些区分是框架性效应(在 Tversky 和 Kahneman 的意义上),它表明了本该是恒定的各情境(的描述)之间的那些差别吗?参见 Amos Tversky and Daniel Kahneman, "Judgment under Uncertainty: Heuristics and Biases", Science 185 (1974): 1124 - 1131; 重刊于 *Judgment Under Uncertainty*, ed. Daniel Kahneman, Paul Slovic, and Amos Tversky (Cambridge: Cambridge Univ. Press, 1982)。导致/允许的区分与一个底线(baseline)之间的关系,不是很可疑地像获得/失去这个区分与其底线的关系一样吗?当然,后者对框架效应来说是一个有利的例子。

许其发生;(5) 不阻止也不避免其发生(当你本来可采取行动时);(6) 不帮助或者不推动那些阻止它发生的行动(当你本可以采取行动时);(7) 不帮助或者不推动它的不发生(当你不能采取任何措施时)。

第二个向度同样涉及行为的因果作用与效果、结果或引致的事态之间的关系,但是这一向度标示的不是一个行为的重要性而是相关于该结果的因果**强度**(robustness)。这个思想是:当某物被作为一个目标来追求时,那么有些假设对于该行为人就是成立的。为了达到那一目标(或者为了更有可能达到那个目标),他会重组其行为。在稍微不同的环境下,他的**这个**行为无法达到该目标,他会反过来做一些其他**将**(would)达到该目标的行为;他会排除掉那些不可能达到该目标的可选行为。另一方面,当某事只是其行动的一个可知的副效果,且如果事实证明他(当前或计划)的行为不会产生这个附带的效果时,他也不会改变自己的行为。当然,如果这个行为反过来会产生另外一种他想避免的严重后果时,那他或许会改变行为。这是一个行为在其中会改变的情境**范围**的问题。相对于明知某事是追求另一目标所附带的效果而行为时的情况而言,把某事作为目标来追求包含一个更广的情境范围的假设。这两者之间还夹有一种情形,即目的在于把某事仅仅作为实现另一个目标的手段。这里在某些情境下为实现该手段而重组行为——不同于附带效果的情况——只是相对于该效果就是一个目的或目标本身来说,可能的情境范围要狭窄一些(例如,考虑这样的情境,那里特定的效果不再是作为实现该目标的手段起作用)。

在因果作用强度这一向度上,我们能区分出个人及一个行为与一个效果或结果之间的(至少)六种联系。行为能够:(1) 旨在作为一个目的的效果;(2) 旨在只作为手段的效果。

或者它能够：(3) 目的根本不在于效果。在目的不在于效果的行动中，该人也许：(3′)知道该效果(并不是目的所在)；(4′) 不知道他本该知道的那种效果(并不是目的所在)；(5′) 不知道有这种效果(并不是目的所在)，而且也不是她本该知道的情形。或者(6′) 该事态(并非目的所在)的发生纯属偶然。

使用这两个向度以及它们的分类，我们可以组成一个 7×6 矩阵。[如果这两个向度不是完全独立的，那么有些矩阵条目(boxes)就是不可能的。]行动和个人与其效果之间的关系会由它在矩阵之中的位置(亦即根据在两个向度上的位置)而得到确定。[①] 行为和结果之间的因果关联模式可做出更精细的区分，决策理论对此是否应当加以考虑呢？如果是，那么要如何考虑？在证据关联和象征关联之内也存在更精细的区分，这应当为决策理论所标出并且加以考虑吗？(后面列在强度向度上且参考知识的分类是否置于证据向度上会更好呢？)我提出这些问题并不是想在此做出解答，而是想把它们列入议程。

62

这两章关于原则和象征意义所讨论的论点也同样适用于伦理原则。当我们把一些行为恰当地归类为一组时，一个行为就代表所有这些行为，而且所有行为的分量都被加于这一个行为，即任何一个行为，且赋予了同等的(象征)负效用。义务性约束可能展示了相同的现象。把行为归类到一个禁止它们发生的原则之下——“不准谋杀”——这会使得此行为脱离对它的成本与收益的孤立效用主义(或利己主义)计算。一个行为变得代表了整类行为，并且承担其分量。这不一定要通过采取一种绝对禁

① 我们可能出于其他的目的而想扩展这样一个矩阵，加上第三维度来代表后果或效果的数量。(Justin Hughs 向我建议在法律语境下以这种方式来扩展矩阵。在这种语境下，我们也许想知道效果有多坏：目的所在的那个后果有多坏，所发生的后果又有多坏。)但是，在决策理论中，这个数量当然已经被结果的效用所表达了。

止的方式(无论如何都禁止该行为),但是通过把它大大增加了的(象征)负效用投进任何计算中,这都给做这些行为设置了一个大得多的阻碍。①

现在我们回顾第一章关于遵守伦理原则之象征意义的讨论。伦理行为能够象征(并表达)作为一个为自己立法的理性造物,作为一个目的王国中的立法成员,作为一个价值与个性的平等源泉和承认者,等等。这些由该行为所象征地表达且例示出的伟大事物的效用,遂被整合进了该行为的象征效用,并因此进入了该行为的决策价值。因此,这些象征意义成为个人做出伦理行为的一部分理由。若个人最大化其行为效用,即最大化其决策价值(DV),他就可能被引致去做出伦理行为。这个人将是在追求**自己的**目标(这未必是**自私的**目标)。根据阿玛蒂亚·森的分类,这个人从事的就是对自我-目标的追求,而不是举足轻重地从事追求**她**自己的总体目标的活动。② 但请注意,如果进入这一更深入的分类(即举足轻重地追求自己的总体个人目标)本身对于行为人具有象征效用,那么这就将进入他的 DV 之中。这样的话,当纳入这种象征效用而行事时,他不是再次在追求他自己的目标,亦即他的修改过的 DV,以致他(在 DV 框架之内)想进入森的另一种范畴的企图注定要失败吗?无论我们对这一点的看法如何,还是存在着更一般的意义。做一个有道德的人,是象征我们最重视之物的最有效方式之一,而且这也是一个**理性的**人所不希望放弃的东西。 63

我们在第一章中探讨了原则的各种各样的功能。我们看

① 回顾上面对于一个不违反任何原则的元原则的讨论,这可能使得违反任何原则代表了对所有原则的违反,由此给予了每个原则以过重的义务论分量。

② 参见 Amartya Sen, *Ethics and Economics* (Oxford: Basil Blackwell, 1987), pp. 80-88。

到，接受并遵守一项特定的原则被看作是一个(一般)行为 A，并且得以在很大程度上是工具性决策理论的框架之中处理。如今我们已经提出了决策理论的一种替代性框架，不仅仅是只包含因果工具性，它还包含了证据方面和象征方面。在这个框架之内，接受一项原则这个行为将具有一种决策价值 DV，而且当它的决策价值最大时，人们就会(从各种备选行为之中)选择它。我们为什么要具有原则——如果具有了那个 DV 原则，我们为什么还要具有其他的原则——我们为什么会有这些具体的原则，这一更宽泛的框架开放了一条对它们做出修正讨论的大道。

3. 合理信念

信念在什么时候是合理的？我们为什么希望信念是合理的呢？我们如何判断它们是不是呢？我们能够如何改善它们的合理性程度呢？

哲学文献中充斥着这样两个论点。第一，合理性是个理由问题。信念是否合理要取决于持有该信念的各种理由。这些是我们认为信念为真(true)(或者也许是认为它具有其他某种可欲的认知美德，比如说解释力)的理由。第二，合理性是个可靠性(reliability)问题。产生(且维持)信念的过程或程序能够高比例地导致真信念吗？合理信念就是由那种能可靠地产生真(或者具有其他可欲的认知美德)的信念的过程而产生的信念。

这两个论点单独都无法穷尽我们的合理性观念。没有可靠性的理由看来是空洞的，而没有理由的可靠性看来是盲目的。它们一起构成了一个强有力的联合，但它们是如何确切地相关的，且为什么呢？

我们自然地把合理性视为一个目标导向(goal-directed)的过程。(这既适用于行动的合理性，也适用于信念的合理性。)传统社会中的行为陈规在于，人们采取某种方式的行动是因为一直以来就是那样做的。相对比而言，合理性行为旨在实现人们具有的各种目标、欲望和目的。基于这种工具观，合理性就在于有效地(effectively)、有效率地(efficiently)实现各种目标、目的

和欲望。对于目标本身，工具观几乎无话可说。① 如果合理程序就是那些能够可靠地实现确定目标的程序，那么当行为是由这样的程序所产生时，它就是合理的，当个人恰当地运用合理程序时，他就是理性的。

65 行动合理性的这种论说参考了产生出该行为的过程或程序。然而，决策理论提出的行动合理性的标准论说却只参考了行为的最大化期望效用。行为也许实现了目标或最大化了期望效用，但可能并不是合理地实现的。这可能是瞎猫逮着了死耗子，无心插柳柳成荫，也可能是系列错误互相抵消了。那么，(给定行为者的目标)该行动本来就是要做的最好的事情，只是它可能不是合理地做的。

因此，看来若决策理论要成为一种合理性理论，那么它也必须要参考产生行为的程序或过程。期望效用(或 DV)公式标出最好的行动。当行为由倾向于产生最好行为(即具有最大的 DV)的过程所产生出来的时候，它才是合理的行为。(的确，决策理论考虑也会进入对过程的最优评估中。)决策理论本身是一种最好行为的理论，而不是合理行为的理论。当考虑应用决策理论时，我们往往会忽略这一点。因为这样产生的一个行为不就是被一个能可靠地创造最优行为的过程所产生的吗？而这样做了的话，它不就是合理的行为吗？话虽如此，但这只是一个**经验性**主张。决策理论堪定最好的行为，但是我们也许是该理论不可靠的应用者：我们也许倾向于忽略掉某些因素，在计算中

---

① "理性(reason)具有一种完全清楚和确切的意义。它指示了选择那种能够达到你想要实现的目的的正确手段。它与如何选择目的完全无关。" Bertrand Russell, *Human Society in Ethics and Politics* (London: Allen and Unwin, 1954), p. viii. "理性(reason)完全是工具性的。它不能告诉我们往哪儿走；它最多只能告诉我们如何到那个地方去。它是杆可开火的枪，可用来服务于我们所追求的任何目标，不论好坏。" Herbert Simon, *Reason in Human Affairs* (Stanford: Stanford Univ. Press, 1983), pp. 7–8.

出错,或者会以其他方式误用这个理论。也许另一种理论在实际上会以高得多的比例产生出最好的行为,即使正是决策理论提供了判断何为最好行为的准则(criterion)。①

信念的合理性也能以这种工具性的术语加以思考。在这种情况下,各种具体目标就是确定的:例如真理(truth)、避免谬误和解释力等。让我们暂时假设只有一个认知目标:相信真理。那么,合理信念就是通过"能可靠地产生出真信念"的过程所产生出来的信念。这一过程不仅包括获得一种新信念,还包括维持、放弃或者修正某种现存信念的方法。② 尽管绝大部分情况下,我将只着重于获致新的信念。

人们一直提出"可靠性在何种语境下成立"这样的问题。程序是只需在被实际应用时才是可靠的,还是在现实世界中所有能应用的时段中都要是可靠的,抑或在类似于现实世界的世界里也要是可靠的,还是一般而言在我们相信的那个世界里(无论我们的这种信念是否是正确的)都要是可靠的?③ 这里还有合

---

① 若实现做最好行为这一目标的最有效过程并不涉及任何我们能够称之为"程序"的东西,亦即不包含任何对于各步骤的有意识监控或者是应用任何规则、原则或理由的话,情况又会如何呢?

② 对于处理这个问题的不同进路,参见 Isaac Levi, *The Enterprise of Knowledge* (Cambridge, Mass.: M. I. T. Press, 1980); Gilbert Harman, *Change in View* (Cambridge, Mass.: M. I. T. Press, 1986); and Peter Gardenfors, *Knowledge in Flux* (Cambridge, Mass.: M. I. T. Press, 1988)。

③ 参见 Alvin Goldman, *Epistemology and Cognition* (Cambridge, Mass.: Harvard Univ. Press, 1986), pp. 58 - 121, 他提供的不是合理性的而是证成的可靠性分析;Frank Ramsey, "Reasonable Degrees of Belief", 载于他的 *The Foundations of Mathematics and Other Logic Essays* (London: Routledge and Kegan Paul, 1931), pp. 199 - 203; William Talbott, *The Reliability of the Cognitive Mechanism* (Ph. D. diss., Harvard Univ., 1976; rpt.,有一个新序言, New York: Garland Press, 1990); Stephen Stich, *The Fragmentation of Reason* (Cambridge, Mass.: M. I. T. Press, 1990) pp. 89 - 100。

66 理程序必须具有何种可靠性程度的问题。它必须是可得到的程序中最可靠的吗？还是从一个相当可靠但不是最好的程序中产生的信念也是合理的呢？程序必须追求更大可靠性吗？或者即便某一程序的可靠性小于 50%，但只要它是能够得出正确解释的最可靠程序，那么由它产生的解释某种现象的信念仍然是合理的吗？在评价一个程序时，我们除了要看这个程序产生出正确结果的次数比例外，不是还要看这个程序产生出错误结果时，会发生什么情况，会导致多坏的情境吗？正如合理的行为并不总是那个最可能达到可欲结果的那个行为一样，我们要使用的合理程序也并不总是那个最可能实现目标的那个程序。决策理论的常见考虑在这里就进来了。①

---

① 可靠性观点的现代根源是 Charles Peirce，他根据"推理的前提为真，其结论为真"的次数百分比来谈论推理规则（主要原则）的有效性。尽管有效的归纳规则会显示一个非常高的百分比，但演绎有效规则的得分是百分百。参见 Charles S. Peirce, "The Fixation of Belief", 载于他的 *Collected Papers*, (Cambridge, Mass.: Harvard University Press, 1931－1958), 5: 223－247，重刊于 *The Philosophy of Pierce: Selected Writings*, ed. J. Buchler (New York: Harcourt, Brace, 1950), pp. 5－22。要注意，在 Peirce 意义上的高百分比并不足以证成可应用那个推理规则。假定在通常情况下，当一个 p 类叙述为真时，另一个对应的 q 类叙述也为真。（我们可以将此表述为有关百分比或者概率的统计陈述。）这意味着我能够可靠地从 p 推出 q 并且指望我那次（大概）在正确的百分比内吗？不是的。情形也许是当我相信一个 p 类陈述时，相关的 q 类陈述通常不为真。毕竟，我相信 p 类陈述的那些场合并不是当 p 为真的情境的随机样本，也许因为我得到证据的过程具有偏见或我本身具有偏见，这些场合是不具代表性的样本。即使一个这种形式统计陈述为真，即"通常情况下当我相信一个 p 类叙述时，一个 q 类陈述为真"，情形仍然可能是，当我从一个我所相信的 p 来推出 q 时，q 通常是假的。因为我从 p 推导 q（或基于 p 作出任何推导）的那些次数不一定是我相信 p 的次数的随机样本或代表性样本。"在统计上，通常当我相信 p 且我从 p 推导 q 时，那么 q 为真"，我们是否反而要这样说呢？而且正是这一陈述许可了对 q 的推导吗？但这是按照何种推导规则推导的呢？我们是否需要在推理规则自身之内规定，这是一个代表性情形，其中 p 类陈述是真的、为人所相信，且 p 用作依据该规则进行推理的一个基础。对一个基于普遍为真（universally true）的主要原则的推理，我们无须担忧推理的具体场合。但（转下页）

这两个论点(理由与可靠性)看来能够很容易关联起来：如果你出于支持性理由而持有信念，那么你会更多地得到真信念。通过赋予“理由”以这种主要作用，我们使得形成信念的过程是可靠的。 67

但是，什么会使得某一事物成为一个理由呢？出于理由而相信为什么能够有助于持有真信念呢？相信的目标为什么要是真理(truth)呢？当个人的信念是为了(可靠地)达到其他目标(例如使他高兴或者使别人非常喜欢他)的过程所形成的，那么他是不理性的吗？我们能找到一些精确的规则或者条件，从而确定何种信念是合理的，何种信念是不合理的吗？

(接上页) 是对于一个基于统计原则的推理，我们担忧推理的**这个**场合是不是代表性场合。

合理性的可靠性分析背后的根源观念是合理性首先适用于一个过程或程序，然后派生地适用于作为该程序的一种例示(instance)的某一特定的推理、信念或行动。这种例示从其所例示的程序那里获得其合理性。当该程序的可欲特性是其可靠性，即能高比例地产生出真信念时，那么该程序就很可能产生出真信念。这是否告诉我们，这个过程所产生的**某个特定信念**很可能是真的呢？那个信念或推理可以归入许多可能的程序之下，它还可以归入许多其他的类别之下。给定有关某组信念的一般概率信息，从某信念是该类成员而推出该信念具有某个特定概率为真——在这种情形下，这些信念所属的那种类型是由一个特定的程序所产生的——这样的推理是一种(概率文献所称为的)**直接推理**。由于特定的例示能够归入许多不同的参照组，所以构建一种据以判断直接推理的有效性准则是一个很微妙的问题，这个直接推理对一个具体情形产生了一个[可分离的(detachable)]概率判断。参见 C. G. Hempel, “Inductive Inconsistencies”, 载于他的 *Aspects of Scientific Explanation* (New York: Free Press, 1965), pp. 53-79; Henry Kyburg, “Randomness and the Right Reference Class”, *Journal of Philosophy* 74 (1977): 501-521; Issac Levi, *The Enterprise of Knowledge* (Cambridge, Mass.: M. I. T. Press, 1980), chs. 12, 16。对信念或推理合理性的可靠性论说而言，这些微妙问题不仅关涉到为某一具体类型的推理构建正确的原则；它们还影响了合理性概念本身。因为具体实例的合理性乃是根据从一组信念的概率信息做直接推理来加以定义的。

认知目标

若信念是由能可靠且高效地获得特定目标的程序所得出(和维持)才是合理的,那么那些目标是什么呢?据说那些目标通常是**认知**目标——相信真理(truth)、避免谬误或一种更宽泛的混合,即包括解释力、可检验性和理论成效。[1] 认知目标这一概念本身并不是那么界线分明的。真理,是的,还有解释力、理论成效以及范围。但是简明性是一种认知目标吗?计算简单呢?对上帝本性的神秘洞见是认知还是非认知的个人目标呢?当这点在东方理论中来设想时,哪个是开化的呢?

哲学家们探讨的主要认知目标是真理。为什么真理是一种目标呢?一种回答是:真理或者说相信真理,乃是有内在价值的。所有真理都有吗?我发现自己对于一些问题(比如说,关于一些国家的首都在哪里的问题)有错误信念,但我毫不在意,而且还有很多真理——世界上每个海滩上各有多少粒沙子的精确陈述——是我根本不想知道的。并不是每个事实都值得了解甚或值得具有真信念的,尽管特定的目的可能使得这种了解具有意义。我认为,有些事情本身就是值得我们去了解的——例如,能够解释一系列事实的深层真理,那个伟大的解释理论(如果存在一个的话),即关于宇宙是如何起源的解释。无疑,我们可能对于某个特定范围的事实产生智性上的好奇。

68

看来可以合理地认为,我们最开始对真理感兴趣乃是立足

---

① 参见 Thomas Kuhn, *The Essential Tension* (Chicago: Univ. of Chicago Press, 1977), pp. 320 - 339; W. V. Quine and Joseph Ullian, *The Web of Belief*, 2d ed. (New York: Random House, 1978), pp. 64 - 82。

于工具性考虑。在应对世界上的危难和机遇时,真理要比错误强,也比没有信念强。[①] 完全准确的真理并无必要,必要的是信念**足够真**,使得据之行动能够产生可欲的结果就可以了。"有用的真理"(serviceable truth)才是我们需要和想要的,而信念要有用的话并不一定要已经是精确地真的。[②] 因此,真理更像罗尔斯所说的"基本善"(a primary good)那样的东西,亦即对**大**范围的——几乎所有的——目的而言都是有用的。因此(几乎)不论我们的具体目的是什么,真理都是可欲的,能够给我们带来好处。[③] 由此,我们之所以欲求真信念并且关注真理,也许是因为真信念对大范围目的而言都是有用的。不过这也就使得我们对于真理的关切是工具性的——至少开始时是这样。[④]

那么,威廉·詹姆斯(William James)说"有用就是真理"是对的吗?我们也许把詹姆斯看作是描述了真理的**价值**,而并非其本质。我们可以把真理设想成支持且解释有用性的那种属性——无论它是什么。如果一种属性支持对不同对象所做表述的有用性,那么这一属性就必定是非常一般且是抽象地表达的。

---

① 我们对真理感兴趣的工具性基础并不在于:人们因为意识到了这是有工具作用的,所以才想要去相信真理。相反它是这样的:由于相信近似真理的那种有用性,所以有些对于信念真理的关注就得到了进化选择,正如对寻求真理的好奇心那样。尽管某些对真理的明确工具性欲望也可能发展出来了。

② 我们相信 p,并且只关心 p 过去是否有用以及 p 是否继续有用,针对这一观点,也许有人主张我们所相信的乃是"p 是近似为真的",而**这**是我们想要确切为真的。但并不是每一种近似在任何语境下都是有用的。所以,或许我们相信的乃是:"p 是有用的",并且正是这点才是我们确切地想要为真的。然而,为什么又认为这点表明我们关注的是真理而不是有用性呢?

③ John Rawls, *A Theory of Justice* (Cambridge, Mass.: Harvard University Press, 1971), pp. 62, 90-95.

④ 在有些博弈论情境下,不知道正确的概率是有益的。参见 Eric Rasmussen, *Games and Information* (Oxford: Basil Blackwell, 1989), p. 116, "Entry Deterrence IV"。

因此，真理的各种理论——符合论、融贯论等等——就是各种解释性假说，即用来支持且解释有用性这种属性之本质的各种猜测。①

69 一般来说，相信错误的陈述会降低个人的真信念的比例；但在某些境况下，错误的信念也许会有助于最大化这种比例。[若他相信此一虚假之事，那么我便告诉他许多他本来不可能得知的事实。或者我会给他上大学的奖学金，他从大学里能学到很多东西。如果他错误地相信他在某一考试中考得很好，那么他将有动力去(成功地)学到很多真的陈述。]他相信这样一个错误的陈述(即遵循生成此种信念的过程)是合理的吗？我们可以通过一个最大化真信念比例的过程来相信某一陈述，尽管这一陈述本身很可能不是真的。如果认知目标就是追求真信念比例的最大化，那么经由一个能有效地服务此目标的程序所产生出来的错误信念就是合理的。但是，如果真理的认知目标被表达为应用于这种具体情形的一种约束——“不许相信任何虚假之事”——那么这一错误信念也就不服务于此一目的，因此是不合理的。②

这个目的不止能够采取这两种形式。阿玛蒂亚·森提出了这样一种结构，在此结构里有两类目标——即行为者本人这次不做某件 T 类事，以及最大化所有人不做 T 类事件的总数——被分别赋予了独立的权重[作为一个还容纳其他目标的“最大化准则”(maximand)的组成部分]。③ 在这个结构内，行为人的认

---

① 如果最终发现不同类型的陈述的有用性是由不同属性所解释的呢？

② 比较目标与边界约束之间的区分，以及我对于“权利效用主义”所作的讨论，参见我的 *Anarchy, State and Utopia* (New York: Basic Books, 1974), pp. 28-33。

③ Amartya Sen, “Rights and Agency”, *Philosophy and Public Affairs* 11 (1982).

知目标既包括这次相信真理(避免相信虚假之事),也包括最大化她的真信念的比例。因此,(至少)存在三种不同的方式来追求认知的真理目的:作为一种边界约束(side constraint)、作为最大化真信念比例的目标和作为此目标与这次避免相信虚假之间的权衡。因此,当信念是由一个能有效(effectively)且有效率地(efficiently)获得真信念这一目标的程序所得出时,还不足以说它就是合理的。应当用何种结构来评价一种程序是否合理呢?还是说存在三种不同但合法的合理性观念,每一种对于不同的目标都是恰当的呢?

我们在"信念伦理(ethics of belief)"这个专题下讨论过一种情境,那里相信真理明显无助于人们的其他重要目标。例如,法庭向一位母亲出示了她儿子犯有重罪的证据,此证据足以说服她之外的一切人,但是一旦她相信此点,她此后就会陷入苦海。[①] 她相信儿子有罪是合理的吗?一种贝叶斯主义的分析(Bayesian analysis)表明,她也许应当达致一种与众不同的结论:她更了解自己的儿子,因此可以合情合理地从一个不同的先验概率开始。但是让我们假定,她那与众不同的先验概率只是基于对儿子的爱和不情愿相信儿子会犯这样的罪。尽管如此,也许还是有人主张她相信儿子无辜是合理的,因为综合考虑,这样认为将最大化她的期望效用——而这不正是合理行为的准则吗?[②] 同样,个人也许考虑相信某种命题对他产生的某种负面的道德影响——比如说,相信不同人种的智力水平存在着先天差别——并基于此不去检验某些证据,不使自己形成某

70

---

① 文学中更多出现的是女性家长和男性小孩子。是母亲更有爱心吗?(儿子是更易于犯罪的。)还是文学里认为女性更容易卷入情感与证据之间的冲突呢?

② 我们已经看到,那是最优行为的准则。所以假定她是经由一个产生最优行为的可靠程序而得到这个信念的。

些特定的信念。(这些负面影响不受这个人的意志所操控,或者是控制起来很难或代价很大)

我们可以区分:(1)“命题 *p*”是要相信的合理的事情;和(2)“相信 *p*”是要做的合理的事情。通过把“相信 *p*”看作是这个母亲可做之事,期望效用论说可以适用于(2);但它似乎并不适用于(1),至少就考虑这位母亲可能具有的**全部**目标而言。因为对(1)——“命题 *p*”是要相信的合理的事情——而言,看来只应该权衡证据考虑。即使证据考虑论说被证明是工具性的,相关的目标也必须只是**认知**目标——而母亲的幸福,无论有多重要,也不是一个认知目标。①

然而,如果认知目标本身具有的是彻底工具性的基础和证成,那么“看透(look through)”这些认知目标而直至它们必须服务的那些终极目的难道不是合理的吗?如果那些认知目标一清二楚无助于终极目的,那么我们不是应当忽略它们而直接地去追求那些终极目的吗?(若是如此,那么“母亲相信儿子是无辜的”这一命题最终就应该是合理的。)那么,“相信何种命题”才是合理的这个概念[即上述概念(1)]反过来可能会要根据**认知**目标来加以**定义**,但是这种做法不就是要该概念削足适履吗?即使认知目标曾经是立足于工具性的,它们不是能慢慢获得其自身的权威,且这种权威甚至会违背那种使得其成为一种目标的那

---

① 关于“信念伦理”,参见 William James, “The Will to Believe”, 载于他的 *The Will to Believe and Other Essays* (Cambridge, Mass.: Harvard Univ. Press, 1979), pp. 13 – 33; Jack Meiland, “What Ought We to Believe”, *American Philosophical Quarterly* 17 (1980), 15 – 24; John Heil, “Believing What One Ought”, *Journal of Philosophy* 80 (1983): 752 – 765。Heil 做出的区分与我的类似,即命题 *p* 是要相信的合理的事,和相信 *p* 则是要去做的合理的事。

种终极目的吗？①

信念的整体观(holistic view)则主张，增加任何一个特定信念都会对整个信念体系产生一种振荡效应。它会修正许多其他的信念，会改变要进入贝叶斯计算的许多假说的先验概率，还会对个人信念的总体解释的统一性(unification)增加新的难题等等。这样一来，仅仅着眼于相信某事而使人获得直接满足的个人效果，就是不明智的。因为引入这样一个信念的长远效果是无法估量的，尤其是若个人为了容纳此种信念而必须改变一般的信念形成程序。即使在个人层面上，也存在着很强的前设，即产生的直接有益效果将无法抵偿由此导致的其他错误信念的振荡效果，以及持有这些信念所带来的进一步的后果。 71

然而无论如何，相信某一陈述而对个人带来的好效果可能是一清二楚的，无论有什么证据，人们还是可能倾向于只管去相信它。为了避免这样一种强烈而突出的特殊诱惑(我们经常遵从甚或只遵从一次也是会很危险的)，个人也许可以采纳一项原则以便只相信真实的东西，亦即只有证据表明其(很可能)为真的东西。根据我们前面提出的"原则如何帮助个人避免诱惑"的论说，若我们采纳原则，则这一次相信一个虚假就变成代表许多次相信虚假，这一次相信真理就代表多次相信真理。由此，相信某一特定真理就得以具有一种不依附于其实际后果的象征效用。于是，认知目标变得具有了一种独立的权威。即使在"忽视这些目标的*局域*后果是更有利的"情境下，情况依然是如此。我在本章后面的"合理性的诸规则"一节会回到"信念伦理"问题。

---

① 参见 Frederick Schauer, *Playing by the Rules* (Oxford: Clarendon Press, 1991)，该书就规则而言对这个问题做了广泛的讨论。即便它们用来促进的那些终极目标将明显得不到推进甚至会受到阻碍，在此情况下，它们依然是具有分量的权威吗？

对理由的回应性

信念的合理性可能来源于产生并且维持它的那个程序，但是并非每一个(可设想的)能得到真信念的有效方法都能够把该信念标明为合理的。如果撞头或者吞服致幻剂是对某个主题具有真信念的一种方式——这里主题并不是该人是否撞了头或者服用了致幻剂——那么该信念本身不是合理的。(然而，个人若是知道此点，那么为了得到真信念而选择撞头就可能是合理的。)信念的合理性与一个在交叠陈述链条中推理、推论和进行证据评估的严密网络相关联。观察可以进入该网络中，但在某个描述层次上这个过程是命题性的。这表明合理性并不简单地就是**任意**一类的工具性，它要求是某种类型的工具性，亦即理由和推理。然后，假定一种特殊的程序是得出真信念的一个可靠方法。如果由这一程序产生出来的行为或信念要想是合理的，那么该程序不单必须包含一个理由与推理网络，这个网络还必须(在部分程度上)是该程序**为什么**是可靠的理由。理由以及推理有助于该程序的可靠性。①

合理性包含的不单是因为**支持性**理由而去做或者相信某事，而且要考虑(某些)**反对性**理由。卡尔·波普尔(Karl Popper)强调了(在科学领域中)搜集反对一种假说或者理论的数据或证据的重要性。他指出，确证(confirmations)是很容易

72

① 合理性的传统论者集中于理由和推理，并未提及可靠过程；有些更晚近的论者集中于可靠性这一方面而不看重推理。如果可靠性过程总是与这些推理相伴而行的话，那么这些专一性就是可理解的，亦即如果唯一可靠的程序都包含那些类型的推理——撞头并不起作用——且那些推理总是可靠的。

找得到的。一种科学理论的标识就在于它排除掉了某些证据或事实;而我们若要检验这种理论,就不要在它更可能是正确的那些场景中,而要在它最容易出错的那些领域之中来寻找材料。[①] 即使我们不认同波普尔认为“不存在任何**支持性**理由”(且认为唯一的支持性理由就是那些没有找到反对性理由的诸报告)这个观点,这也是一个有益的强调。合理性包含了既要考虑支持性理由,**也要**考虑反对性理由。信念或行动必须不仅仅是(以正确的方式)由支持或反对的理由所引起的;它必须还要**回应**这两种理由。在这些理由的性质、力量或平衡的某种变化范围内,如果理由是不同的,那么行为或信念也将是不同的。[②] 信念或者行为必须是正向回应的:如果理由更强,则该种信念必须不能消失或更弱;如果理由更弱,则此种信念必须不能更强。

哲学期刊的典型文章所具有的结构都是为了在读者当中产生合理信念的。一般说来,一种哲学观点或论点都是作为值得相信的而提出的,且正反理由都得到了考虑。理由中正面的有:它从中推演而来的一般性的和可接受的陈述;它符合或伴随的其他可接受的东西;其可接受的,因此支持它的后果;有它的例示或有契合它的例子,这些也为它提供了一些证据。所考虑的反对论点的理由有:可能的反驳(这些反驳得到了回应,比如被削弱、破坏或者多少被避免了);有可能的反例(它们被削弱了或

---

① 参见 Karl Popper, *The Logic of Scientific Discovery* (New York: Basic Books, 1959), and *Conjectures and Refutations* (New York: Basic Books, 1962), p. 240。

② 我们在此可以提出与知识论中的追踪概念可相提并论的要求,参见我的 *Philosophical Explanations* (Cambridge, Mass.: Harvard Univ. Press, 1981), ch. 3。

者用来把论点修正成了另一个免于此反例的命题,后者则变成了值得相信的合理信念)。在此,有一种反对命题 p 的理由值得单独提起:另一个命题 q,它是 p 的最好的或最可行的替代,可能比 p 更值得我们去相信。做法就是提出对 q 的特定反驳,提出那些认为足够消除 q 或表明为什么不应该接受 q 的困难或反例。极少会出现这样的情况:对 q 的反驳并不比对 p 的强,但有更好的理由支持 p。对 q 做出与对 p 同样的全方位考察就更罕见了。尽管如此,所有这一切还是给该论点增加了正反理由的各方面考量,而使读者有更好的机会来合理地相信它。

或许我们首先就应该说,合理性涉及的是对相关因素的回应,即对所有相关因素也只对相关因素的回应。认为相关的要 73 素是*理由*,则是另一个论点。然而,仍然还有一个更进一步的论点——这个论点可能并非在所有领域都成立——即这些理由可以被整齐地分成*支持*与*反对*两大类别。但是,合理的信念究竟是以何种确切的方式对正反理由进行回应的呢?亦即这些正反理由究竟是以何种方式决定这个信念的可信度(credibility)的呢?个人持有信念是为了最大化那些支持理由的*净*权重,亦即以支持理由的权重量减去反对理由的权重量,这样说太过简单了。正反理由之间可能存在着某种互动:或者支持某事物的理由的权重也许要取决于它们要面对哪些反对的理由;某个东西是*不是*一项支持理由,这甚至要取决于存在着什么样的反对理由。

不仅如此,理由的具体权重可能还取决于那些本身并非正反理由的其他因素。假设某一陈述 r 是相信 s 的一个理由,而 q 是能够削弱 r 的一个反对理由,但不是作为一种理由来认为 r 本身或者 s 是错误的,而是作为一种理由来认为 r(在这种语境

之下)是支持 s 的较弱理由或者根本不是理由。[1] [相应地,也有可能存在着使理由权重增大的因素:加强因子(aggravators)。]因此,支持理由的净权重并不是由这些理由自身单独确定的;其值还要取决于它们周围存在着何种削减因子(和加强因子)。

这提出了正反理由的一种神经网络模式构造。支持 S 陈述的一个理由 r,沿着频道向 S 节点(node)发送一个具有正权重的信号。反对 S 陈述的一个理由 r',沿着频道向 S 节点发送一个具有负权重的信号。支持 S 的理由 r 的削减因子将沿着频道发送一个在 r 与 S 之间频道上减少该权重(可能减至 0)的信号。这一网络的结果便是对于陈述 S 的可信值。(我们将在下一节进一步探讨这种结构。)这一框架可以容纳具有不同权重的多类理由。因此,要么是在这个网络内,要么是作为整个网络的一个突生(emergent)现象,我们也许有望把握来自科学哲学文献中的许多方法论。

尽管理性者关心的乃是真理,他利用具有的或知道的理由净余来估量和预测这种真理。然而,由于理由可能是对真理的一种有偏见的指示器,我们从这些具有的理由净余直接外推来做判断可能是讲不通的。理由据之到达我们的那种过程可能会区别性地允许某类理由,即那些指向陈述的真值而不是指向其虚假的理由。因此,如果把我们的信念只基于理由而不考虑理由的代表性,这是不明智的。 74

这里有个有点人为但不失启发意义的画面。想想个人的理由被看作有关陈述 s 的相关理由总体的一个样本。这个总体可以包括其他人所知悉的事实,包括能弄清楚的事实等(或许有些

---

① 参见 Robert Nozick, "Moral Complications and Moral Structures", *Natural Law Forum* 13 (1968): 1-50; *Readings in Nonmonotonic Reasoning*, ed. Matthew Ginsberg (Los Altos, Calif.: Morgan Kaufmann, 1987); John Pollock, *How to Build a Person* (Cambridge, Mass.: M. I. T. Press, 1989), pp. 124-155。

限制会排除掉 s 陈述自身,即使其他人知道 s,有些限制则会排除蕴含 s 的其他的陈述)。那么,问题就在于那个人的理由是否有偏见,或者在理由总体中是否具有代表性。个人自己可能会有其他的理由认为这种偏见是存在的。[①] 我并不是主张,理性人会首先就用她具有的理由净余去估算所存在的理由净余,然后应用这个净余结果对真理进行判断。但是理性人会尽力去了解,她所具有的那些理由是否是真理的一种有偏见的指示器,并会由此影响她从具有的理由对真理所做的判断。如果她断定自己的理由在特定方向上不具代表性,那么她就会更正它。当用合理性来评价理由时,与此相关的不仅仅是理由的力度,还有理由的代表性。

正是在这个意义上讲,合理性包含某种程度的自觉性(self-consciousness)。不仅理由要接受评价,而且获取、储存以及调出信息的那些程序也要接受评价。理性人将会尽力提防在这些过程中的偏见,并且会采取措施纠正那些已知的偏见。在评价

---

① 但是,如果我们把最后这个理由纳入到我们的正反理由当中去,那么我们能够进行直线外推(straight-line extrapolation)吗?然而,最后这个理由不是直接地作为有关 x 的某种结论的正反理由来关注 x;因为它反过来关注我们获取信息的过程,所以最好在另一个层次上来处理它。

Bernard Williams 认为与个人相关的行动理由只是那些他已有的理由,或者是从他现有的愿望、偏好以及评价出发(如果能获得更充分的信息的话)经由周全的思虑而能够获得的理由。参见 Bernard Williams, "Internal and External Reasons", 重刊于他的 *Moral Luck* (Cambridge: Cambridge Univ. Press, 1981), pp. 101 - 113。这些内在理由与行为人的现有动机相关联。但是当个人问自己,"我应当做什么?"的时候,他不一定是在询问:什么东西能够最好地服务于他现有的动机或他若有更多信息将具有的那些动机。他可能知道其他人的动机不同且在某些方面优于他自己的动机。可能由于一个特别穷困或者受虐待的童年时期,他自己没能发展出那些动机,或者某个特殊的情境给他灌输了某些动机。他想知道的就是,什么才是**最好的**理由,因此他在此点上可以赋予其他人的意见以某种分量。(这已经假定了一种特殊的既有的内在动机,亦即去做最好的理由所支持的事情,Williams 会这样说吗?)

已有信息的重要性时，给定各种各样的事实，个人会考虑他本来可能获得何种不同的信息，获得这类信息的可能性有多大。这是“三囚徒问题”①②给我们的一个教训，此外也是对于鲁道夫· 75

---

① 三个人为死罪而受审且被判有罪，但其中只有一个人被判处了死刑，三人都不知道谁被判了死刑。那么我们可以认为，他们也同样如此认为，每人都有 1/3 的概率被处死。到了行刑前一夜，囚犯 A 请求看守(看守知道哪个囚犯将会被处死)把 A 写给妻子的一张便条转交给另外两个囚犯当中不会被处死的那一个。当看守走开去转交便条的时候，囚犯 A 相信自己有 1/3 的概率在第二天早上被处死。当看守走回来之后诚实地告知他已经把便条交了出去，此时囚犯 A 仍然相信自己有 1/3 的概率被处死。他没有得到任何相关的新信息，因为他事先就知道另外两个人当中(至少)有一个人可免一死，因此看守是能够把便条交出去的。现在他又问看守把便条交给了哪个人，看守告诉他便条已经给了囚徒 B。如果囚犯 A 推理出自己现在有 1/2 的概率被处死，根据在于他与囚犯 C 开始各自均有 1/3 的处死概率，且情境仍然是对称的，因此他们还是有相同的概率，因此上升到 1/2 了，那么他就搞错了。情况并不是对称的。囚犯 B 和 C 都可能成为便条的接受者——所以便条没有被交给 C 这一事实，对于确定 C 的处死概率是相关的(它现在已升至 2/3)——然而 A 本来就不能成为便条的接受者。这个信息本来有可能推得 C 不会被处决，但是这不能推得 A 不会被处决。因此，当实际的信息得到时，是 C 的概率增加了。相对照而言，在以下情境中囚徒 A 的概率将升至 1/2：囚徒 A 向那个走回来的看守询问道：“囚徒 B 收到了便条，对还是错?”然后看守回答说：“对；”或者如果囚犯 A 首先请求看守把那张便条交给一个第二天不会被处死的人，三个人之中的任何一个，然后(假定看守把便条交给两个不会被处死者的可能性是一样的，包括 A 本人)看守回来时说他已把便条交给了囚徒 B。这里的关键因素是，B 收到便条的方式在概率上是不同的，这取决于囚徒 A 或是囚徒 C 是否要被处死的情况。如果囚徒 A 将会被处死，那么 B 收到便条的概率就是 1/2(且 C 收到便条的概率也是 1/2)。如果将会被处死的人是囚徒 C，则 B 收到便条的概率就会是 1。这些概率值相加得到的总和是一又二分之一，其中来源于 A 被处死这一情况的 B 收到便条之概率(即 1/2)占总和概率的 1/3，而来源于 C 被处死这一前提的 B 收到便条之概率(即 1)占总和概率的 2/3。因此，得知 B 收到了便条这一信息，我们就可以推知 A 有 1/3 的概率被处死，而 C 有 2/3 的概率被处死。把上述这些过程放到一个贝叶斯网络表中就可以看得很清楚。见 Judea Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausibile Inference* (San Mateo, Calif.: Morgan Kaufmann, 1989), fig. 9.1, p. 417。

② 在三囚徒问题中得出错误结论的人们——即剩余的两个囚犯各有 1/2 概率被处死——乃是由于适用了某种对他们而言似乎是自明的规则或原则(转下页)

卡尔纳普(Rudolf Carnap)计划的一个额外的阻碍，这个计划是要建立一种归纳逻辑来规定证据 e(一般来讲)在多大程度上能够确证假说 h。[①] 如果个人的信息来源是这样的，以致即使 p 是虚假的，它们也不会(或不太可能会)传输这一信息，那么便不能从他没有得到这类信息做出 p 为真的结论。他必须考虑到他的信息源乃是偏向于承认"p"的[②]。我随后将对理由偏见这个

(接上页)所致，亦即这个规则本身不够精巧，或者他们没有精巧地应用这个规则。我们应当认识到自己也有可能处于这种情境中，即便我们具有的是目前能说出的最好规则，因此不要急于用一种规则来压倒某一可陈述的理由(例如，"你已经知道了他们当中的一个人不会被处死，因此……")。这里也有空间以 Amos Tversky 和 Daniel Kahneman 的风格来探讨，从而确定那些为三囚徒困境所误导的人们会使用何种具体的一般启发法。参见 Tversky and Kahneman, "Judgment under Uncertainty: Heuristics and Biases", 重刊于 *Judgment under Uncertainty: Heuristics and Biases*, ed. Daniel Kahneman, Paul Slovic, and Amos Tversky (Cambridge: Cambridge Univ. Press, 1982), pp. 3–20, 也参见他们在该卷中的其他文章。

① 重要的不仅仅是 e 的具体内容，还有本来还能够得到的其他信息(以及概率有多大)。而试图把后一种信息整合进证据本身中去，就会完全改变一种卡尔纳普式的归纳逻辑的结构。参见 Rudolf Carnap, *The Logical Foundations of Probability* (Chicago: Univ. of Chicago Press, 1950), and *The Continuum of Inductive Methods* (Chicago: Univ. of Chicago Press, 1952)。Carnap 在第一本书所偏爱的体系中包含了一个仔细陈述的无差别原则版本：所有的结构描述都获得了相同的先验概率。也许在最根本的层次上，先验概率的任何差别都必定存在着某种结构性理由(即使这里，也许有人会把此看作一个经验支持的问题)。但是，即使在当前的基础物理学中，也没有理由认为我们已经达到了那一层次，而且我们通常所考虑的那些属性则肯定不是如此。

② 考虑你的信息来源中有一条信息 I 说"p 是真的"。那么收到信息 I 表明了什么呢？作为贝叶斯定理的一个应用，*prob*(p/收到 I)=[*prob*(收到 I/p)×初始的 *prob*(p)]/[*prob*(收到 I/p)×初始的 *prob*(p)+ *prob*(收到 I/非 p)×初始的 *prob*(非 p)]。让我们进一步假定，信息源将报告 p 要么为真要么不为真。(考虑"它们可能不做任何报告"这种可能性会使我们接下来的陈述复杂化，而这是没有必要的。)那么 *prob*(收到 I/非 p)=1−*prob*(信息源说"p 是假的"/非 p)。那么上述贝叶斯定理的规定中的分母就等于是 *prob*(收到 I/p)×初始的 *prob*(p)+[(1−*prob*(信息源说"p 是假的"/非 p))×初始的 *prob*(非 p)]，分母中的后一部分，(转下页)

问题做更多的讨论。

合理性的诸规则

哲学家历来寻求为合理信念、演绎结论的合理推理和接受以及得出信念的(非演绎)合理方式构建规则。他们寻求那些具有表面吸引力的规则(这些规则据其内容向理性毛遂自荐,且产生了我们最确信的推导和信念),亦即可以用来明确(sharpen)我们得出信念和评价信念的各种方法的那些规则。① 然而,如果信念合理性乃是产生并且维持信念之过程的有效性的一个函数,那就不可能保证最优过程会应用任何具有表面吸引力的规则。反而,这些过程会涉及一种在对立规则和程序之间的计分竞争,它们的竞争力取决于(通过特定的计分程序)每项规则在过去成功地参与预测和推理的纪录。规则和程序没有必要在表面上

(接上页)即加号之后的部分,就等于初始的 *prob*(非 *p*) − *prob*(信息源说"*p* 是假的"/非 *p*) × 初始的 *prob*(非 *p*)。因此,给定非 *p* 的情况下,这些信息源说"*p* 为假"的概率越小,那么收到信息 *I* 说 *p* 为真对于 *p* 是真的这一假说的支持就越小。贝叶斯定理告诉我们,我们也必须同时考虑本来还会得到其他什么样的信息,以及以何种(条件)概率得到。

现在我们考虑这样一个贝叶斯分析,即认知的怀疑论假说 *SK* 在面临我们的观察与经验 *E* 时命运如何。既然怀疑论假说已经确定,所以 *prob*(*E*/*SK*)=1, 这得出,即使 *prob*(*E*/非 *SK*)也等于 1, *prob*(非 *SK*/*E*)也不会升到超过非 *SK* 的先验概率。根据后验概率的贝叶斯论说,怀疑论的后验概率不比先验概率小,非怀疑论者的后验概率不会比它的先验概率大。证据是没有用的。

① 讨论原则和事例之间"反思平衡"的文献设定了原则本身具有一种独立的、盛行的表面(on their face)权威,参见 Nelson Goodman, *Fact, Fiction and Forecast* (Cambridge, Mass.: Harvard Univ. Press, 1955), ch. 4, sec. 2, pp. 62 − 66; and Rawls, *A Theory of Justice*, pp. 19 − 21, 48 − 51。对反思平衡的一个批判性讨论,请见 Stich, *The Fragmentation of Reason*, pp. 83 − 89。

看起来是有道理的，但是，在接受反馈而不断地修正后，它们是互动配合来产生出符合所要求的外在准则（比如说真理）的结果。由此，那个过程理论将不会是这样的几个规则，可以看出它们的表面道理，以致个人能够可行地应用它们，而是一个模拟那个过程本身的计算机程序。[①] 更极端一点，即使是在一个权重竞争中，情况有可能是所有规则都不是用符号表达的，而任何"规则"都是作为一个平行分布处理系统（parallel distributed processing system）的行为规则性而出现的。在这个系统中，决定是否激活输出向量或者中间向量的权重矩阵要受到某种纠错规则的反复修正。[②] 如果达到那些认知目标的最有效过程便是这种类型，那么哲学家们试图构建的那类规范性规则就将无法胜任为信念合理性划界。这些规则本身不会成为那个过程的成分，而且自觉地应用它们也并不是达到真的（或者可欲的）信念的最佳路径。

76

如果是这样，那么在得出信念的可靠过程的研究中，哲学家们在技术上将变得过时。他们将被认知科学家与计算机科学家、人工智能的研究者和其他人所取代。[③] 我们的理解将得到

---

① 关于这类理论的一个框架，参见 John Holland, Keith Holyoak, Richard Nisbett, and Pual Thagard 令人着迷的书，*Induction: Processes of Inference, Learning and Discovery* (Cambridge, Mass.: M. I. T. Press, 1986)。

② 参见 *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, ed. James McClelland and David Rumelhart, 2. Vols. (Cambridge, Mass.: M. I. T. Press, 1986)，特别是 chs. 1-8, 11, 14, 26。

③ 已经有数量庞大的文献标明了这种转向。关于一般性的反思，参见 Clark Glymour, "Artificial Intelligence Is Philosophy"，载于 *Aspects of Artificial Intelligence*, ed. James Fetzer (Dordrecht: Kluwer, 1988), pp. 195-207。我的重点并不在于划分学科边界。当然，现在那些正在接受哲学训练的人，可能会把工作转移到这些领域里，而在将来，已经进入到哲学学科里的一些人，也会进入到其他那些领域里面去。有趣的是关注已经改变的理论目标的性质，还有随之而改变的那类理解的性质。尽管如此，哲学家们对于各领域和各种任务之中包含什么东西的概念分析工作，可能仍然有助于人工智能和认知科学的研究者避免那种不可能成功的研究路径。

进步,但是理解的性质会发生改变:电脑模拟将取代(理论提出的)揭示结构(structurally revealing)的规则,这种规则具有表面有效性,人们能够理解和应用它们。① 这对我们是有用的——机器将会被制造出来以完成复杂的任务——但这并不是哲学家们早先所期盼的:能够用来改善我们自己信念的那些规则和程序,亦即**可察知的**(surveyable)规则和程序(我用的这个术语来自维特根斯坦),我们可把它们吸纳进来作为一个整体而加以理解,由此对合理性的本质能提供一种揭示结构的描述。(思考下面两者之间的区别:一是传统的下棋策略建议,另一个是通过某种反馈性计分程序且**只**根据过去步骤的结果而学习的程序。后种程序化机器可能会下得很好,只是我们不能理解它在任意一场具体的游戏当中是如何做到的;从中我们学不到任何可以遵循的规则来改善自己的技巧。程序所得出的精细权重系统可能无法用我们具有的概念表达出来。)然而,这些程序所栖居的机器可能会变成改善我们信念有用的外部助手。(我可以相信计算器对某一问题的回答,而无须去理解它为什么是正确的答案。) 77

假设达致信念的最可靠程序涉及这种计分程序和权重的不断修正;仅有的规则便是决定什么东西以何种竞争力进入到竞争中,决定那些竞争力如何受到获胜方的确定结果所修正。②[这里有一项规则:如果你射高了,那就稍稍调低一点与正数互动的正权重;如果你射低了,那就把与正数互动的正权重稍稍调

① 现如今,模拟(simulations)在物理学与社会科学中是很普通的。但是就我所知,若科学产生的不是一个理论陈述或一般法则体系,而是一个程序和一个模拟,科学哲学家还没有考虑过解释理论所出现的这些特殊问题。

② "但是,如果这个装置的判定与我们对某一具体情形或命题的直觉相左的话,我们还会信任它吗?"我信任的肯定不是我自己对于这步棋的最好选择,而是这一非常有经验的下棋机器的建议,它的下法是按照纠错规则来修正权重的结果,即便我自己搞不懂其建议的任何理据。

高一点;不断地这样调整,直到你准确地射中目标为止。无疑,这是一种令人赞赏的规则模式——具体的规定将精确地说出要对权重做出何种调整——但这不是哲学家们过去一直寻求的那类合理信念的原则。“桶队列算法 (bucket brigade algorithm)”也不是分配信用(apportionment of credit)的,[①]尽管它处理的东西比德尔塔规则(delta rule)处理的东西更像规则式实体(rulelike entities)。]即便如此,人们也许会说,哲学家的原则具有一种揭示作用,亦即,能够描述这些反馈程序(依据过去实际预测与推导中的表现来调整权重的计分程序)的**输出**。有些原则可以**定义**认知目标,并因而能够确定程序所指向的目标,但是这些(尚有的)原则可能无法更有揭示力地进一步描述那一输出。我们无法事先就知道,各种哲学原则——这些原则不只是定义认知目标——是否能够准确地描述能最有效地获得那些目标的程序的结果。

例如,思考经常提出的这个规范性要求,即个人的信念体系应该是前后一致且演绎闭合的(deductively closed)(所信事物的任何逻辑推论也应该是被相信的)。也许,达致高比例真理(和错误相对更少)的最有效程序将会产生出一组不一致的信念。故而我们若想要维持高比例的真理,则信念集最好不要是演绎闭合的。这样,如果事先就要求不得使用任何已知(在某些环境下)会导致不一致信念的规则来产生信念,那么这可能会阻碍我们获得大量的真信念。尽管当这类规则得到适用时,我们必须采取措施限制那些可能会出现的不一致后果。我们可以寻

---

① 参见 John Holland, *Adaptation in Natural and Artificial Systems* (1975); rpt, Cambridge, Mass.: M. I. T. Press, 1992), pp. 176 - 179; Holland, Holyoak, Nisbett, and Thagard, *Induction*, pp. 70 - 75, 116 - 117。

找一种可接受的方法来避免不一致,但与此同时我们要进行损害控制,要隔离不一致,采取措施防止从明确的矛盾当中推出任何一种武断论述。① 在日常生活中,我们很平静地这样做——我们乐于承认我们自己的易错性,认可我们有某个信念无疑是错误的。科学体现了维持一致性的一种强大功力,但是科学家们也在等待时机,现在还是不会放弃掉那些众所周知会导致不一致或者不可能值(impossible values)的理论的诸多精准预测[见证了在量子论的计量中对重整化(renormalization)的使用]。 78

尽管如此,哲学家构建合理信念的明确原则这一传统努力仍然还有一种作用——一种谦和得多的作用。一项明确的具体原则 P 的判定,即认为某一陈述 q 是合理的(或者是不合理的),可能是一种信念过程的结果之一。原则 P 的这一判定将不会是信念独自的决定物,而是会进入信念形成过程的后续阶段,且在未来的信念形成中,将根据我们所观察到的接受还是拒绝陈述 q 而行动的结果来修正其权重。这样的一项规则或原则 P 可能是一个平行分布处理系统之中的一个处理单元,而它的输出——不管是合理的判定,还是一种贝叶斯概率,或者是其他的什么东西——可能会扩展到后续阶段,起到激活下一个处理单元的作用,直到得出最终的结论(无论是否是信念),且其进一步的结论会得到确认。② 然后,赋予原则 P 与整个综合系统里其他成分的关联权重,将会按照某种学习规则,根据最终结果的 79

① 关于把集合在理论悖论上和语义悖论上的矛盾隔离出来,并且限制它们的危害的讨论,请见 Ludwig Wittgenstein, *Remarks on the Foundations of Mathematics* (Oxford: Basil Blackwell, 1956), II 80 - 82, III 60, V 8 - 12。

② 平行分布处理系统的倡导者们倾向于认为:规则是作为改变权重分配的一种结果而出现的,其自身不必在任何场合被象征性地表达——我倾向于说,规则乃是经过一种看不见的手的过程而出现的。这里的"规则"是体现于诸多(转下页)

反馈而被修正。(或许原则 P 的某些细节也会得到修正，它由此而转变成了一个新的原则 P'。)

我这里并不是要认同平行分布处理模型的支持者们所采取的(严格的)研究程序，我的论述并不依附于这些程序，尽管这种关联主义(connectionism)理论在当下很是盛行。① 稍后我将假定特定规则会把一种结果传输到一个复杂的反馈程序之中，并且我根本不在意这些规则自身是作为平行分布处理系统中的一种规则性，还是可以用符号明确地表达出来的。对于我的目的而言，平行分布处理系统的启发之处在于它所提出的一般性框架：多重单元按照由某种反馈(纠错)规则不断予以修正的一个权重矩阵而进入进一步的诸单元(其活动取决于有什么东西进入它们)。这样一种网络的有些单元自身可能就是**规则**。但是

(接上页)处理单元之间的一种关联性模式(pattern of connectivity)。其他还有可能找到规则的地方是：每个处理单元的输出函数，通过网络传播活动模式的传播规则，把影响某一处理单元的输入与该单元的现状相组合以产生一个新的激活水平的激活原则，关联模式受到经验修正的学习规则或纠错规则等。[此处，我遵循了平行分布处理模型中所列的清单，参见 D. E. Rumelhart, G. E. Hinton, and J. L. McClelland, "A General Framework for Parallel Distributed Processing"，载于 *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, *Rumelhartand McClelland* 编(Cambridge, Mass.: M. I. T. Press, 1986, p. 46)。]在最后一个方面，亦即学习规则中，我们才预计会发现某种符号表达(symbolic representation)。无论如何，我此处的话语是用来表明，在某种平行分布处理模型的一般性框架内，一项规则是如何具有一种功能的；我并没有做出这一强假设，即规则只能以关联模式表达出来。关于批评不以符号形式提出规则的关联主义心理学理论之适当性的论文，参见 *Connections and Symbols*, Steven Pinker and Jacques Mehler 编 (Cambridge, Mass.: M. I. T. Press, 1988)。

① 参见 Paul Churchland, *A Neurocomputational Perspective* (Cambridge, Mass.: M. I. T. Press, 1989); Andy Clark, *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing* (Cambridge, Mass.: M. I. T. Press, 1989); Patricia Churchland and Terrence Sejnowski, *The Computational Brain* (Cambridge, Mass.: M. I. T. Press, 1992)。

在平行分布处理结构中,它们在自身之间展示出的关系与那些非规则单元之间的关系是类似的。

在这个一般性框架中,我们可以说传统上的哲学家希望构建的是这样一个原则,它通过使该原则的输出成为随后信念具有任何权重的唯一输入而决定该信念状态。而在一个平行分布的处理结构中,诸如原则这种成分,可能不必用上述方式就能取得那种强作用;随着时间的推移,其输出值能影响信念单元的其他那些(独立)成分之权重有可能会变为0。(因此,这个一般框架为传统哲学家的希望留下了空间,但并不依赖它。)然而,即便达不到这么强的作用,一项明晰的(explicit)原则还是有帮助的,即其裁定是达到信念的最可靠过程的一部分。我们对于明晰原则应持一种有保留的态度。

科学哲学文献中提出了各种各样的方法论准则,它们都可以嵌入在这个一般性框架之中。有些准则也许就是系统中的成分,其权重逐渐得到修正;有些准则是作为系统运作的描述而不是其成分而出现的。[①] 我想到的是诸如这样的准则:证据越是多样(这种多样性由体现在你其他信念中的范畴所判断),支持就越大;假说越简单越好;一种理论或者假说继承了其所取代的理论的证据支持力;新现象的预测比对既有现象的推导能增加更多的可信性;要避免特设(ad hoc)假说;控制好对结果很可能产生影响的其他一切变量;产生的预测越准确,对该假说的确证就越大。 80

我们已经看到一项明晰原则如何在产生真信念(或满足某

---

① 在"Simplicity as Fallout"(载于 *How Many Questions: Essays in Honor of Sidney Morgenbesser*, ed. Leigh Cauman [Indianapolis: Hackett, 1983])中,我曾提出过一种观点,即满足简单性准则是如何从一个体系的运作中出现的,而不是作为该体系内的一个评价成分而出现的。我在那里所考虑的体系并不包含符合某种纠错规则的反馈。

个其他可欲认知目标的信念)中起到有益的作用。但是一项明晰的原则能够帮助我们理解合理信念的本质吗?理解合理信念的一种方式是将其看作某类过程——例如,一种平行分布过程(根据明晰原则来说,理解的一种最差情形)——的结果。或者我们可以更详尽地把信念合理性理解为这种具体过程的结论,它具有所有参数、功能以及规定的转换规则。(请比较:现今的动植物生命是由那些特定的达尔文过程持续运作而产生出来的;一个市场社会中的现有财富及收入的分配是由那些特定过程的持续运作而产生出来的。)虽然我们或许还想知道其他的东西:也就是说,当前这些过程的结果所表明的模式,就它们确实展现出了某种可以被描述的模式而言。现今的生物又是如何起作用的呢?收入与财富的分配模式是什么样的,而且它们又与哪些因素相关联呢?合理信念集看起来又是什么样子的呢?如果合理性的目的在于真理,那么,得出合理信念的原则或许就不是对于合理信念是什么样的最佳描述。合理信念是什么样的,对这个问题的最好的简短描述或许就是真理集的最扼要的描述——这个真理集是我们能够简单地描述而无需遭遇对角化问题(diagonalization problem)的最大集。这个集合与合理信念集不是一模一样的——我们常常都会相信虚假,而且我们还不相信有些真理——但这或许是我们所能提供的最简单、最易理解的描述,而且这个集合中存在的那些不匹配,也会通过我们正在提供描述这个事实而减少,因此,在走向我们这个信念集时,我们对于真理集的描述也会包含错误。

尽管我们想要的可能既不是对于我们现时段合理信念的模式的描述,也不是对于当前信念所得以产生出来的那个过程本身的描述,相反,我们想要的是我们现今的合理信念与其来源之间的那种关系的描述,亦即对于我们当前合理信念的内容与结

构是如何与生成它们的那个事物的内容与结构关联到一起的，对此给出一种具有结构阐示力和相对简洁的描述。这种关联的模式又是什么呢？或者根本就不存在任何这样的因素需要理解。由竞争性的计分反馈程序所建立和支撑的这一关系，可能并不展示任何具有结构揭示性的模式，即便只是大略的近似。(这样就不存在我们所不理解的事情了，根本就不存在要理解的这种东西。)但是，如果对于这种关联，或它的一个后继阶段确实存在着一种模式，那么我们就能够用某种明晰原则表达它。这样的一项原则也能帮助我们确定：通过表明他们所说的东西与其所依凭的根据之间的那种关系，其他人告诉我们的事情，我们对之应该给予多大的信任。①②

81

---

① 还存在其他可能的路径来评价另一个人的可靠性，这种可靠性不谈及何种关联模式划定了他的信念，例如，有关可靠性的统计。但是一般来讲，这些路径是很难获得的。然而"为什么你这样想?"对此也许有一种相对简单且更具阐示性的答案。他回答"我对这些事情有丰富的经验，因此我知道"，这种回答也许是有用的。用现在的术语来说，他告诉我们，他当前的权重和回应都是由大量的反馈信息塑造的，且得益于这些反馈信息。只要我们认为这个陈述是系统能够经由学习去认识到其真理的那类陈述，那么这就是令人安心的。

一个过程只要最终能得到某一主题的真理——它最终对情形能够得到高比例的真信念——那么它就是可靠的，即使这需要一段很长的时间。如果在这一过程中，信念也是这样来得到的，即受到反馈信息的修正等等，那么许多早期信念就一直是虚假的了。当前的信念可以是由一个可靠的程序生成——它最终会得到高比例的真理——但如果过程的某一阶段赋予它为真的概率还未超过 1/2，那么现在相信它是合理的吗？一个建议是把这个程序置入某种外部装置之中，直至该装置稳定地搞定了某个信念后才去相信它。(那么，我们的可靠程序就是**那样的**装置了。)然而，我们可能不具有这种选项。或许这种程序只能在我们的头脑中运行，而且反馈只对我们实际相信的东西起作用。由此，要运行这一程序，即最终会可靠的程序，就会要求，在那条并不是很可靠地为真的路上，我们实际上要具有许多信念。(若科学家必须相信当前的结论才能进行常规的科学研究，产生要求用新理论来解释的异常等，那么科学就是这样的一个例子。)

② 有效学习程序之范围与限度的研究，最近得到了正式发展，由此产生出了许多有用的区分(对于一个领域中的每一个陈述而言，会存在某一时刻，该(转下页)

现在，我想描述一下这个网络(它将决定我们的信念)中具有(按照某种反馈规则)可变权重的诸多成分中的一个。我们从贝叶斯定理来开始是有帮助的，这一定理可以轻松地从概率论公理中推导出来。它陈述的是假说 h 基于证据 e 的概率要取决于什么。我们不仅仅可以通过复杂的公式，也可以根据直觉陈述来理解这一定理。

基于某一数据 e，假说 h 有多大可能性呢？数据通过假说 h (中的真值)而产生的可能性又有多大呢？“e 要归因于 h 的成立”这个事实的可能性有多大呢？这不简单地是如果 h 为真，则会产生 e 的那种概率。我们暂时把这个概率记为 $prob(e:h)$。也就是说，给定 h 为真，数据 e 的可能性有多大——这里我宽松点说——*如果 h 为真*，e 有多大的可能性。

为了知道数据 e *确实是*经由假说 h 而产生的可能性有多大，我们必须考虑的不仅仅是 $prob(e:h)$，还有 h 本身有多大的可能性。如果 $prob(h)$ 的值很低的话，那么即使 $prob(e:h)$ 的值很高，e 也不大可能从 h 中产生。如果有一族隐藏的外星生物(Alpha Centurious)具有极大的能力，他们现在最想让我听到号角声，那么在那一刻外面马上有号角声的概率就是极高的。然而后一假说本身的概率是极小的，而且该假说也不会因为有能力解释那一特殊的号角响声而赢得很高的概率。正是 $prob(e:h) \times prob(h)$ 才表达了从真理 h 之中产生出 e 的可能性，亦即从真理 h 产生出 e 有多大的*绝对可能性*。

82

但是，这告诉了我们 h 有多大的可能性或者是 h 基于 e 的

(接上页)程序此刻得出那一陈述为真吗？对于一个领域中的每一个陈述而言，会存在某一时刻，该程序此刻得出所有那些陈述都为真吗？)以及许多有启发性的结论。参见 Daniel Osherson, Michael Stob, and Scott Weinstein, *Systems That Learn* (Cambridge, Mass.: M. I. T. Press, 1986)。

可能性有多大吗？我们知道了 e 通过 h 而出现的绝对可能性，但是 e 也有机会从其他假说 $h2, h3, h4$……当中产生。我们想要看到的不仅仅是 e 从 h 中产生的绝对可能性，还有**相对**可能性。e 能够从所有(其他)方式产生的可能性的多大百分比才是 e 从 h 中产生的可能性呢？(把括号内的“其他”两字删除的话，对于数学演算而言将是更加便利的)e 产生于 h 的**相对**可能性乃是这两者间的比值：即 e 产生于 h 的绝对可能性，与 e 可能产生于所有不同方式的各个绝对可能性之和。看来正是 e 产生于 h 的这种相对可能性告诉我们 h 基于数据 e 的可能性有多大。我们刚才一直考虑的 h 重命名为 $h1$。那么，$h1$ 基于 e 的概率看来就是：

$$\frac{prob(e : h1) \times prob(h1)}{\sum_{i=1}^{n} prob(e : hi) \times prob(hi)}$$

这看来还告诉我们在给定 e 的情况下，$h1$ 有多大的可能。因为它告诉我们，相对于 e 可能从中产生的所有途径而言，e 有多大的可能性是从 $h1$ 中产生。因此，我们似乎具有了对贝叶斯定理的一种直觉推导，或者至少是一种解释。

但这还不完全是我们所得到的那个贝叶斯定理。因为我们所谈的概率，即如果一个假说为真，则它会引起证据 e 的概率 $prob(e : h)$，并不是贝叶斯定理中所指的条件概率，后者不过是给定某假说成立，则该证据为真的概率，而不管该假说是否引起了那一证据，即不管该假说与证据之间是否存在着任何(概率上的)假设性关联。但我们得出的是贝叶斯定理的一种因果版本或说假设版本。为了强调这些假设概率不是标准的条件概率，我们把如果 h 为真，则 e 为真的概率记作：$prob(h \rightarrow e)$。

上述版本多少还需要一些修正，而且不仅仅是印刷上的修正。因为，尽管我们已经考虑到了本可能引起证据 e 的各种不

同假说，但我们尚未考虑到这一可能性：即 e 有可能是自发出现的，而并不存在（概率式）原因或者生成的假设性事实。这点也需要纳入考虑。（此处我们可以把关于 e 的这种机会假说记作 Ce，不过是否认存在着任何 hi，以致 hi 确实概率式地产生了 e。）因此，我们的公式变成了下面的：

$$\text{度量}(h1/e) = \frac{prob(h1 \rightarrow e) \times prob(h1)}{\sum_{i=1}^{n} prob(hi \rightarrow e) \times prob(hi) + prob(Ce)}$$

这个度量测量的到底是什么呢？[①] 这个公式是否把 $h1$ 基于 e 的测量确定为一种条件概率，亦即解释为一种条件赌博中的下注赔率呢？还是这个公式确立了某种其他因素的数值，比如说 e 支持 $h1$ 的程度，或者是如果 e 为真，则 $h1$ 为真的概率，亦即 $prob(e \rightarrow h1)$ 呢？我们要探讨的一个预备性问题是，公式右边的这个比例的属性是什么？它是否像概率那样运作，诸如此类？

有一篇讨论解释推理（explanatory inferences）问题的哲学论文，名为“最优解释推理”。[②] 说存在这样一种推理的原则就

---

① 要注意这个测量避开了“已知证据问题”。参见 Clark Glymour, *Theory and Evidence* (Princeton: Princeton Univ. Press, 1980), pp. 85 - 93; Daniel Garber, “Old Evidence and Logical Omniscience in Bayesian Confirmation Theory”, 载于 *Testing Scientific Theories*, ed. John Earman (Minneapolis: Univ. of Minnesota Press, 1983), pp. 99 - 131; Colin Howson and Peter Urbach, *Scientific Reasoning: The Bayesian Approach* (LaSalle, Ill.: Open Court, 1989), pp. 270 - 275; John Earman, *Bayes or Bust* (Cambridge, Mass.: M. I. T. Press, 1992), ch. 5。因为即使证据 e 是已知的，并且 e 对 $h1$ 的条件概率为 1，尽管如此，[$prob(h1 \rightarrow e)$]（即如果 $h1$ 为真，则它产生 e 的概率）仍然不必等于 1。如果它等于 1，那么，因为 $h1$ 通过使其全概率进入分子而可使得它在这种测量中的值增大。

② 参见 Gilbert Harman, “The Inference to the Best Explanation”, *Philosophical Review* 70 (1965): 88 - 95; Norwood Russell Hanson, *Patterns of Discovery* (Cambridge: Cambridge Univ. Press, 1958), pp. 85 - 92。

是主张 h 作为一种解释性假说的好性质(goodness)足以决定 h 作为一种信念的可信度。尽管更弱的主张看来才是真的：也就是说，h 作为一种解释性假说的好性质只是确定及评价 h 之可信度的因素之一。[①] 上面的比例在规定和分析这一因素时看来是比较有用的。h 作为 e 的解释到底有多好，(至少)要依赖于如果 h 为真，则 e 为真的概率，且还要依赖于 h 的先验概率；对这些因素的依赖看起来对它们的乘积(product)会有影响。其他情况相同，如果 h 的乘积大于 hi 得出的乘积，那么，假说 h 就比 hi 提供了更好的解释。

然而，我们并非总是想要推出那个最优解释。设想有 8 种可能的解释 hi，每种假说 hi 的 $prob(hi \rightarrow e)$ 都相同，而且 $prob(h1)$ 只是稍大于 1/8，而其他 $h2$，……，$h8$ 的概率每一个都稍小于 1/8。在这种情况下，$h1$ 可以是最优的解释，但是尽管如此它并不是个很好的假说；这是由我们方程式的右边所测出的 $h1$ 的小比例所标明的。当然，一个假说以前的成功解释是很重要的，这些将影响到该假说进入到公式时的先验概率。其他解释特征也能够影响到这个先验概率。因此，也许我们的这个比例式测量的是，e 对于 $h1$ 的解释支持度。[②] 在评价一个假说的解释性地位(explanatory status)时，相关的考虑不仅仅是有待解释的单个事实 e：何种假说从"有待解释的总体事实"中获得 84

① 或许我们应该把最优解释推理的倡导者理解成把它作为一个可挫败的推理原则而提出来的。

② 也就是说，给定假说已经被构建出来且已知的情况下，e 给予 $h1$ 的解释性支持度。某个其他的尚未被构建出的假说 $hn+1$，也有可能解释 e，而且这一假说并没有包括在公式分母的任何一个因素中，也没包括在 e 产生的偶然概率这一最后的因素中。因此，这个比例式测量的是：相对于我们现在已构建的可容许的各种解释性假说而言，e 给予 $h1$ 的解释性支持度。

了最大的解释支持呢?[①] 在一个以可变权重向前注入的多成分网络(例如一个平行分布处理网络)中,h 的解释值将只是进入到 h 的总体可信度之中的一个因素。

让我们设想一种网络纳入了很多因素——包含贝叶斯概率、解释值(如因果贝叶斯公式所表达的)、波普尔式方法论准则(Popperian methodological maxims),还有对削弱因素(undercuttings)的评估——这些因素向前注入而得到陈述 h 的一种**可信值**。我们把此视作对 h 的正反理由做出恰当权衡的一种理想评价。[②] 我的直觉是,在这个体系中,内容丰富的假说会变得越来越丰富。新的数据不会算作是对每一个与之相符的假说的证据,它们会被假说中最可信的那个所占有,并用来增强该假说的可信度,而不会被用来增强其他假说的可信度。(这一点可以例示于这样的体系中,那里现存的假说竞相争取新数据的支持;而那些已经具有较高可信度的假说在这种竞争中具有一种优势。)当过去最可信的那种假说出于某种理由被拒绝的时候,那么支持它的那些数据便可以为另一种假说所利用。[③]

---

① 为评价一组事实对于某一假说的解释性支持度,我们是应当采用这些事实的合取且把它看作是我们因果化的贝叶斯公式中的证据 e,还是当这些事实在逻辑上独立时,我们先一个个地处理它们,即一次评价一个事实的因果化贝叶斯值,然后再把这些值加起来呢?

② 关于贝叶斯条件概率的网络理论,请见 Pearl, *Probabilistic Reasoning in Intelligent Systems*。

③ 比较 Charles Peirce 的"惯常法(method of tenacity)",("The Fixation of Belief",载于 *The Philosophy of Peirce*, ed. Justus Buchler [London: Routledge and Kegan Paul, 1940], pp. 5 - 22), Nelson Goodman 的"加固(Entrenchment)"概念(*Fact, Fiction, and Forecast*, pp. 87 - 120, and Robert Schwartz, Isrel Scheffler, and Nelson Goodman, "An Improvement in the Theory of Projectability", *Journal of Philosophy* 67 [1970]: 605 - 608),以及竞拍系统,见于 Holland, Holyoak, Nisbett, and Thagard, *Induction*, pp. 70 - 78, 116 - 121。

这种处理系统是一种学习系统;其权重经由反馈而加以修正。但这种反馈又从何而来呢?该体系中的预测和期望的纠正可以是源于外部老师输入的纠正值,或者是缘于意料外的感觉输入,再或者是源于该体系所记录的会降低每个参与成分之力度的内部不和谐。① 具体的纠错规则可能会随体系的不同而不同,在一个体系内部也会随任务的不同而不同,甚至纠错规则自身可能也与一个遵循进一步的纠错规则的反馈体系内的纠错规则相竞争,或甚至就是这些竞争者之一。

个人若被看作这样一种体系,也会表现出几个与评估其总体合理性相关的方面。她有一套好的权重和强度体系吗?她有一个监测体系来记录适当的反馈信息吗?她有一个能有效率地修正权重的纠错规则吗?以及(或许)她有对整个网络的结构进行修正的某种程序吗?这整个体系要成为合理的,并不要求任何一方面达到最优。或许还存着更有效率的纠正规则,它能够更迅速地收敛到不需纠正的恰当权重,但是如果现有的纠错规则至少能够在正确的方向上和用正确的标记来做出(小小的)修正,那么整个体系就是(或多或少)合理的。我们需要思考的是,在体系作为一个整体而起作用的过程中,各组成部分是如何相互关联的。 85

至此,我们已经设想了一个体系,它给每一个待评估的陈述 h 计一个分数,由此能够生成各种可信值。那么,所得出的这个 h 的可信值又是如何被用来得出有关 h 的信念呢?这种接受规则又是什么呢?

第一条接受规则是这样的:

规则 1:如果与 h 不相容的其他陈述比 h 的可信值更高,则

---

① 参见 Holland, Holyoak, Nisbett, and Thagard, *Induction*, p. 9。

不要相信 h。

可信值与概率不同,因为 h 的可信值与非 h 的可信值之和并不一定等于任何固定值。然而,如果非 h 的可信值比 h 的更高,那么规则 1 告诉我们不要相信 h。(但是,既然有许多因素,诸如解释力,都参与了对可信值的确定过程,所以不要假定相对来说不确定的非 h 总是比 h 有更高的可信值。)

请注意,此规则还适用于这种情况:与 h 不相容的那个备选陈述不是在那个层次上的某一种可选假说,而是 h 的某个反证 g,它与 h 严格不相容——一种波普尔式证伪者(falsifier)。在这种情形下,如果 g 的可信值高于 h,那么我们就不应该接受 h。但是这种明显的否证(disconfirmation)也不是结论性的,因为 g 的可信值可能低于 h,或者也有可能随着 h 逐渐增大 g 的某个削弱因素的权重而使 g 的可信值下降。[①]

规则 1 排除一些陈述,给我们当前的信念留下一组可选项,亦即其可信值未被与之不相容的某种陈述所超过的那些陈述。在这些可容许的陈述之中,我们又应该相信哪一个呢?我们是应当全部相信,还是经过某种去除不相谐的进一步修剪程序后,去相信余下的一切呢?我认为我们不应当遵循信念的这种最大化策略。在那点上,我们相反应该遵循一种持有某一信念之可欲性的决策理论计算(decision-theoretic calculation),亦即除了认知目标之外,还包括实践目的与效用的计算。(我不是说我们应当明确地进行这种计算;而是说我们的行事应该符合它,一旦

---

① 我们可以把认识怀疑论者看作是提出了一条更为严格的规则:如果一种陈述的可信值没有达到它本能达到的值,那就不要相信它。但是这一规则可能会有不同的理解:如果该陈述的可信值低于其他任何一个陈述之可信值的话,而不论该陈述是否与第一个相冲突;如果它的可信值能为某个其他的证据或理由所提高的话;或者某个陈述具有更高可信值在逻辑上是可能的话。

对这种计算的偏离引起了我们的注意,那么就要修正。)这就是我们的第二条接受规则。

规则 2: 仅当相信 h 不比对 h 不具任何信念的期望效用低时才相信(可容许的)h。① 86

我们由此具有一个两阶段程序:第一阶段排除具有较小可信值的陈述;第二阶段,通过考虑这种信念(广义的)的结果而在余下的陈述中确定我们的信念。这个程序很值得推荐。在这种程序之下,个人就不会持有认知上(根据可信值判断)低劣的信念。然而,他也不是必须对该问题具有信念。他是否具有信念要取决于他的目的和目标,既有实践和理论上的,也有社会的。

我们再来考虑那个重罪犯母亲的例子。我们假定对她而言,儿子有罪比无辜的可信值要高——即便她更了解自己的儿子,这一点实际上可能无法颠倒其他人得出的可信值。然而,相信儿子是有罪的,即使这种信念是准确的,还是会对母亲产生很大的负效用。按照这里提出的两阶段原则,相信儿子是无辜的对她来说仍是不合理的。(儿子是无辜的,这不是一件她要相信的合理的事——规则 1 就排除掉了那种信念。)然而,她不相信儿子是有罪的,这却不是一件不合理的事情;在这个问题上不持有确定的信念,对她来说并不是不合理的。(但是若她不相信儿子是无辜的,这将肯定会给她带来痛苦,那么情形又将如何呢?此处我们可以援引前面所作过的区分。儿子的无辜不是她要去

---

① 一个更严格的规则要求,相信 h 要比不对 h 具有任何信念能获得更大的期望效用。二者的区别就在于,相信 h 与不对 h 有任何信念的期望效用持平时,我们是否该相信 h? 我们或许觉得应该相信,因为相信真理具有一种价值——但根据假设,这种价值不是已经包括在产生平局的那个效用计算中了吗? 或者我们可能认为应选择不信 h,因为持有信念这个行为本身是有成本的(我们只有有限的记忆能力、精神能量或者其他的什么东西)——但是,同样,这些成本不是也已经包括在效用计算当中了吗?

相信的一件合理的事；但是"相信那点"或许是她要做的最好的，因此，也是要做的合理的事情。）但是假如有罪的证据是**压倒性的**又如何呢？但在我们相信特定的某件事之前，一定要满足的那个标准之严格程度在部分程度上是由我们决定的。我们决定何种证据水平，即何种程度的可信值，才能说服我们；而且我们把这种水平定得有多高可以取决于诸多因素，包括持有某些信念对于我们和社会的效用。（那位母亲不需要把标准设得无法满足。也许她能决定，只有当她直接且不可抗拒地体验到上帝直接告诉她，她儿子有罪时，她才相信这一点。）我们能够选择确立标准有多高，但这并不意味着我们可以随心所欲。也许一致性的原则或要求会包含：人们看作类似的情形应该满足相同的标准；标准的严格程度须直接随着情形在相关向度上（例如，基于信念的效用量）的位置变动而（且永远不是反向）变化，如此等等。①

87

要注意到，只要一个陈述的可信值小于某个与它不相容的陈述的可信值，那么在第一个阶段它就会被排除在信念的候选之外。②然而，一个信念的候选并不一定要具有最大的总体可信

---

① 法庭内的旁听者可能会告诉那位母亲："我们认为证据充分，不相信您的儿子有罪将是**不合理的**。"那位母亲和旁观者们都同意，可信度较低的陈述（即那位儿子是无辜的）不会被相信，但他们的分歧在于何种可信度必然使得相信他有罪。那位母亲可能会问，他们如何决定在何处设定证据充分性的那个门槛。看起来可行的是，若这个门槛的设定是他们按照那个门槛指定持有某类信念的普通效果所决定的，那么这位母亲不是能反驳说，对她而言这个效果不是普通的效果吗？而且如果**他们的**普通效果和她实际上具有的那个效果是一样大的话，那么他们本来也会设立一个更严格的门槛。既然如此，那么为什么她的信念受制于只是适应于他们的不同情境的那种标准的指令呢？（毕竟，陪审团成员也会恰当地在有别于普通旁观者的地方设定他们的门槛。）

② 因果贝叶斯定理在评估某一假说的解释性支持度时，也会考虑（分母中的）所有的备选假说。然后这一结果也会作为一个因素进入评价该假说之可信值。由此规则 1 要求我们进行比较的不是与该假说不相容的所有假说，而是其中具有最大可信值的那个。当且仅当该假说过了这一关之后，它才能够作为信念的候选项。

值。它的可信值可以小于其他备选的但只要不是不相容的陈述的可信值。还要注意到,尽管一个陈述在第二阶段上被拒绝作为信念——出于实践理由相信它是不可欲的——但它在第一个阶段上依然能有效地排除那些与之不相容且可信值更低的陈述。① 在这种原则下,信念成为了理论考虑与实践考虑的一个组合,其中理论考虑(词典式)优先于实践考虑。

第一个阶段是消除性阶段,这一点给人这种感觉:就信念合理性而言,"不合理性"乃是首要的观念——用奥斯汀(J. L. Austin)(性别主义的?)的话说就是:是穿裤衩(当家做主)的那个。显而易见的是,对很多事情的相信是不合理的。但不那么显而易见的是,某些信念是如此可信以至于是合理性所指令(mandated)要相信的,也就是当你根本未对这个问题持有任何信念的情况下,你不相信它们就是不合理的。在某些环境下,个人可以比其他人施加更严格的信念标准,并因而在他人持有某种信念的场合下他根本不具任何信念,这并非是不合理的。 88

规则 2 并不指令一种信念作为合理的。如果一个陈述过了第一关——没有不相容的陈述具有更高的可信度——然后,规则 2 告诉我们,如果相信该陈述比对这个问题不具任何信念所获得的期望效用还要小,那就不要去相信它。因此,规则 2 告诉

① 假设 p 比 q 有更高的可信值,而 q 又比 r 的可信值更高;再假设 p 与 q 不相容,q 与 r 不相容,但 p 与 r 是相容的。那么,r 是因为 q 和应用第一规则而成为一个不被容许的候选项,还是 r 是可容许的,因为本来会排除它的 q 本身将会因为 p 和第一规则的应用而是不可容许的呢?依据第一规则而排除某一陈述的东西本身必须是能见容于第一规则的吗?在发展这个系统时,探讨所有这些方向将是有益的。

情境或语境的每个变化都必然要使我们重新计算所持有的每个信念(以及非信念)的期望效用,第二规则会这样指令吗?这里把获得一种信念与保持或改变一个信念区别开是有用的。这里有一个惰性规则,即除非有某种特殊的理由要求改变,否则不变。

我们什么时候不要去相信一个陈述,而没有告诉我们要在何时去相信它。

但如果一个陈述果真通过了规则 1 的检验,并且如果(这个人)相信这一陈述比对该问题不持任何信念有更大的效用,那么他不相信这一陈述难道不是不合理的吗?(在此之前,我一直忽略了这一可能性:即两种不相容的陈述可能在可信值上是平手,即彼此都不占优。这样,每个陈述都留下来作为信念的可能候选项。在这种情境下,决策理论计算就必须不仅要比较相信每个陈述的效用与在该问题上完全不具任何信念的效用,而且还要比较相信这个陈述的效用与相信另外一个陈述的效用。假如比较之后再次是平手,那么,也许两个都可以。)不具有信念将一无所获,且还会损失某些收益——这是实践计算的决定。如果一种陈述有足够的可信度——也就是说,不比任何不相容陈述的可信度低——决策理论计算推荐人们去相信它,个人**难道不该**这样做吗?我们可以把规则 2 的(更)严格的版本转换为信念的一种充分条件:

规则 2′:如果相信 h 的期望效用比对 h 不具任何信念的期望效用更高,那就相信(可容许的)h。

然而,若对可信值如何起作用没有一个更详尽的理解,那我是不愿接受这一规则的。一种陈述的可信度比任何不相容陈述的可信度都要高,但其可信度仍然很低,存在这种可能性吗?[不像概率那样,互斥(exclusive)且穷尽(exhaustive)陈述之可信度之和不一定等于 1。这样的话,一个陈述可以比它的否定的可信度要高,但它的可信度仍然很低,难道不可以出现这种情形吗?]在那种情形下,就不存在要相信它的指令(人们没有对该问题持有一种信念的迫切需要)。

这对于合理的信念提出了另一个要求:仅当一个陈述的可

信度足够高时才相信。[①] 至于多高才是足够高,这将随着陈述种类的不同而不同,比如它是一种观察报告、理论科学的陈述,或是对于既往历史事件的一种信念,等等。由此,

规则 3:给定一个(可容许的)h 陈述的类型,仅当其可信值足够高时才相信它。

什么因素会来设定每类陈述的可信度水平呢?是能使持有该类信念(按照规则 3 持有信念的情况下)的效用最大化的那种水平吗?那么,就出现了两种效用计算方式:第一种是根据规则 2 的指令对某一个具体信念的效用计算,第二种是在规则 3 中用来为某一类信念设立可信度水平的效用计算。但是,既然陈述可以用各种方式进行分类,那又是什么决定了相关类型的规定和形式呢?除非对什么算做一个类别存在着某种限定,否则这种效用计算就有使规则 3 蜕变为规则 2 的危险。[②]

一项陈述通过这三项规则的检验——假设规则 3 能够被充分地确定——确实为我们提供了合理信念的一种充分条件。由此,我们又有了规则 4:如果一个陈述没有为前三个规则所排除,那么相信它。(我们已经得出:仅当一个陈述没有被前三项规则中的任一项所排除时,才相信它。)更明晰的表述是:

规则 4:当没有任何一个与 h 不相容的备选陈述比 h 的可信值更高,且给定 h 的陈述类型,它具有足够高的可信值;并且相信 h 的期望效用至少与不对 h 有任何信念的期望效用一样大的话,那么相信 h。

---

① 如果进一步的探究表明,对于规则 2′的这种担忧是毫无依据的,那么,这个补充要求就是没有必要的。

② 比较规则效用主义蜕变为行动效用主义的那种危险。

（再次有一个更严格的规则要求：相信 h 的期望效用比不对 h 有任何信念的期望效用要大。）规则 2 是用决策理论的标准框架表述的，谈及的是期望值。如果我们前面的讨论是正确的，那么决策理论计算就应当最大化行动的决策价值，即其因果效用、证据效用还有象征效用的加权总和。因此，第二规则可以更恰当地表述如下：

规则 5：仅当相信（可容许的）h 的决策价值至少与不对 h 具有任何信念的决策价值一样大时才相信 h。

为了决定相信哪种陈述，我们不仅要考虑持有一种信念的因果后果，而且还要考虑这种信念作为证据而指出的东西，还有这种信念的象征效用。（规则 4 应依据谈及决策价值而不是期望效用加以重新表述。）

我早前说过，一种生成信念的规则毋须保证所产生的所有信念都是一致的，甚至在某些环境下可以论证会导致不一致的信念——不过在那种情形下，我们应该采取措施隔离并限制那种不一致的后果。想想克伊布格（Henry Kyburg）那个得到广泛讨论的"抽奖悖论"例子。[①] 举行一次抽奖活动，你知道这一百万张奖券中有一张奖券会胜出。对于任何一张给定的奖券而言，它不被抽到的概率是压倒性地高的。如果一个陈述充分高的概率是成为一种（合理的）信念的充分条件，那么对于每一张奖券而言，你都会相信它不会被抽到；因此你相信奖券 1 不会被抽到，你相信奖券 2 不会被抽到……你相信第一百万张奖券也不会被抽到。但是，你确实相信抽奖会举行且会有某一张奖券

90

---

① 参见 Henry Kyburg Jr. *Probability and the Logic of Rational Belief* (Middletown, Conn.: Wesleyan Univ. Press, 1961), pp. 196 - 199, and "Conjunctivitis", 载于 *Induction, Acceptance, and Rational Belief*, ed. Marshall Swain (Dordrecht: Reidel, 1970), pp. 55 - 82。

被抽到。因此你相信：或者是奖券 1，或者是奖券 2……或者是第一百万张奖券会被抽到。故而这整个信念集是不一致的。有些哲学家得出：这个例子表明，若一种规则仅仅因为某事物的概率超出了某一很高的值便推荐相信它，则这种规则是不恰当的。不仅如此，如果该规则继续说，一个人仅当某事物的概率超出那个值才相信它，那么，他不会总是相信他所相信的两种事物的合取，因为该合取的概率比任何一个不确定的合取的概率都要小，并因而会在某些情形下降低到信念的那个临界概率以下。

我们所提出的观点并没有使得概率成为陈述可信值的唯一决定因素。然而，我们的观点确实面临着这类问题：只要我们相信两个陈述的每一个，那么我们也要相信这两者的合取吗？这里提出的这种结构表明肯定会导致不一致性吗？而且，如果是这样的话，我们将援用什么规则来限制它们的影响呢？（我们的讨论可以依据规则 2′而不是依据那个更为复杂的规则 4 来进行；任何关于信念的充分条件都将面临类似的问题。）

应用规则 1（连同关于平局的那个条款）的人不会去相信她明知会不一致的两个陈述。尽管如此，由于“认出不一致性”并不是一个机械问题，所以无法确保一个真诚的规则应用者不会被（不知不觉地）导向不一致的信念。尽管一旦这种不一致性被发现，规则 1 就会认为我们可信值更低的那个陈述不能被作为信念的备选项。因为规则 1 聚焦的是与一种陈述不一致的其他备选陈述，所以规则 1 的精确应用能够保证个人的信念是成对地一致的。这个人不会同时持有彼此不一致的两种信念 $h1$ 和 $h2$。但是，更大的不一致圈子又如何呢？在抽奖悖论中，我们有一百万张奖券中有一张会赢的陈述；同时我们还有每一张奖券

都不会赢的单个陈述。在这一百万零一个陈述中,任何两个陈述都是符合成对一致规则的——两个陈述都可以同时为真——但不能所有一百万零一个陈述都为真。这些就构成了一个不一致的集合。

然而,我们并没有要求整个信念集必须是一致的,我们仅仅要求它们必须是成对一致的。如果你想要一个高比例真信念的话,那么你应当相信这一百万零一个陈述中的每一个陈述;则你将有一百万次都是对的。有人可能会反对说:"但是如果这个集合是不一致的话,那么你知道你必定会错一次。"确实如此,但是,如果换种情形来说,有这样一种形成信念的过程,即我知道它每一百万零一次便会给我生产出一个错误的信念,那么不作为一个逻辑问题——这些信念全都是一致的——而是作为一个事实问题,我选择让这个过程来形成我的信念难道不是合理的吗?如果这个错误是由一个逻辑问题而产生的,因为这个信念集是不一致的,那么事情会怎样变化呢?亦即遵循这一信念形成过程的可欲性会发生怎样的变化呢?的确,当我们知道这些信念不一致时,为了避免不一致性,我们最好确保自己不要再把所有这些信念都用作一个论证的前提;但这是隔离不一致后果的问题。

91

在这个情境中,我们不能无限地组合信念来得出新的信念。知道某一特定的陈述(比如说一个合取)是不一致的,这点会(在决定可信值的要素网络中)进入到对该陈述所导致的可信值中,从而使该陈述获得一个最小的可信值。(当然,这也取决于"所引入的合取是不一致的"那个特定陈述的得分。)对一种已知为不一致合取的否定将与该合取不相容,因此总是能获得一个更高的可信度分数。因此,根据规则1,那个不一致的合取不能作为信念的一个可容许的候选项。

但是，我们这种不一致的合取可以走多远呢？我们可以把多少个一致的陈述联合起来并且相信它呢？在抽奖情境中，规则1（和关于平局的规则）阻止我们走向两个不同且彼此不一致的合取中的任何一个——例如，这一陈述：有一张奖券会赢但不是前五十万张奖券中的任何一张；以及另一陈述：有一张奖券会赢但不是后五十万张奖券中的任何一张。不仅如此，在这些抽奖的情形中，我们联合得越多，所得到的合取的可信值也就越低。（这种概率降低也是影响所得可信度的要素之一。）在某个关键的集结点上，很可能它的可信度会降到低于与之不相容的某个相竞争陈述的可信度，并因此不容许作为信念的候选项。如果就可信度而言，这个情境是对称的，那么我们就会得到许多结构相似的合取，按照规则1来说，每个合取都是信念的备选项。但是，规则2′无需认可每一种这样的最大合取，即便不考虑平局的情况。真有什么报偿是靠相信这种最大的（一致的）合取而得到的吗？（别忘了与这个最大合取相一致的某个较小的合取仍然会有更高的可信度。）当我们在抽奖的例子中应用规则1和规则2′，并且以可信度（而不是简单的概率）来操作时，我们就可以预期下面的东西。

个人会相信这张或那张奖券会中。然而，指定任一奖券，他认为它不会中。对于任何给定的两张奖券，他认为没一张会中。对于任何成对奖券而言，他相信也会双双落选。对于任何三元组也是类似的。这种情况会在某一点上变得模糊起来。对一个很大的 n 元组（n-tuple），他并不持有全组内没有一张奖券会中选的那个信念。他不再相信前90万张奖券中无一命中。他甚至不相信前499 999张奖券中会无一命中。那么他的信念会止步于何处呢？这有什么重要性吗？可能在某个具体情境中的一个点上，他会相信对于这种或那种的大量合取，且在该情形中， 92

他是否持有那个信念由持有它的效用所决定。现在，那个论述——至少直到前一句话——听起来就非常像我在抽奖问题上的情境了。[①]

如果我们的信念可以是不一致的——尽管规则致力于使它们保持成对一致——我们如何隔离这种不一致性也许会带来的那种危害呢？众所周知，使用标准的逻辑推演规则，我们能从不一致性当中推演出所有陈述。[从 p& 非 p 中，我们可以推出 p。从 p 我们可以推出"p 或 q"(q 为任意一个陈述)。从 p& 非 p 中，我们可以推出非 p。从"p 或 q"以及非 p 中，我们可以推出 q。不仅如此，我们不必从这个明显的矛盾合取"p& 非 p"开始；可以仅仅从这两个明晰陈述开始，即陈述 p 和陈述非 p，其后的推理大体同前。]

人们可以采用多种方法来避免这种信念的升级。我认为，信念要在演绎推理中合法地从前提转移到结论的话，必须不仅相信每一个前提，而且还要相信这些前提的合取。(或至少，保证这些前提所组成的合取不是不可信的。)这给予我们在适用演绎推理中对于信念的第六项规则。

规则 6：仅当每一个逻辑前提 pi 都是可信的，**并且**只有当这些逻辑前提的合取 $p1$& $p2$&……& pn 也是可信的情况下，才可以因为 q 是从前提 $p1$，……，pn 经由明晰的演绎推理而得出的，从而相信 q。

当我们把一个陈述作为信念的一个备选项评估时，规则 1 吩咐我们要确定是否存在另外一种陈述，它与此陈述不相容但

---

① 要把合取的可接受规模的模糊性消除到比决策理论计算所做的更为精确，我并不觉得这是一个很迫切的问题。人们在抽奖悖论问题上的兴趣点并不在于它的**连锁诡辩**(sorites)方面：探讨究竟多少粒沙子组成了一堆沙子？

又具有更高的可信值。这里不存在这样的假设,即每个可容许的备选项都必须通过一道统一的可信度门槛。陈述 $s1$ 可能被另一个不相容且具有更高可信值的陈述 $s2$ 排除在信念候选项之外,但是陈述 $s3$ 与 $s1$ 不相关但比 $s1$ 的可信值更小,不过它也许会是可容许的信念候选项,因为没有一个与它不相容且有更高可信值的陈述。可信度的门槛是一种比较性门槛;因此,当不同的语境涉及具有不同可信度的不同类型的竞争陈述时,门槛就会随语境的不同而不同。

93

规则 2 和规则 5 认为:相信一个可容许陈述的后果要决定是否相信这个陈述。任何一种完整的理论都会允许"相信 p 蕴含什么"的某种考虑。因为相信 p(出于某些理由或是作为某种程序的结果)就将成为这个世界中的另一个事实,如果这发生的话,那么它也会有其自身的意义。各种陈述的条件概率可以发生变化,包括我们关注的这个陈述 p 本身的条件概率;给定支持 p 的各种理由,并且给定你也是因为这些理由而相信 p,p 的条件概率可能就不同于单独基于那些理由的 p 的条件概率。或者对 p 的相信可能会产生一些因果后果,而这会在随后的情境中改变什么东西为真。①

虽然规则 5 比这走得更远,它增加了相信 p 的因果效用、证据效用和象征效用的考虑,但这些效用并不能使一个不容许的假说成为合理信念的可容许的候选项。(另一方面,尽管 p 并不是要相信的合理的事,但规则 5 并未否认在某些环境下,相信 p 仍然是一件要做的合理的事,亦即,相信一个不容许的 p 也许

---

① 参见 Richard Foley, "Evidence and Reasons for Belief", *Analysis* 51, no. 2 (1991): 98-102; Richard Jeffrey, "The Logic of Decision Defended", *Synthese* 48 (1981): 473-492。

会有最大的期望效用或决策价值。)①

有些人会避免在一个给定的社会中探究某些主题,因为预见到它产生的结论——某个真信念也会有很多的扭曲和误用——会对社会造成有害的后果。同样,某些人也避免去相信某事,因为他预测到了这种信念实际上将会给他自己,亦即给他的性格以及行为模式带来实际影响。但这并不要求他持有一种相反的信念,只是要求不持有这种信念而已。同理,因为一种信念指明的(即使不是引起的)有关某人的东西,且因为这些信念所代表和象征的东西,他会避免具有这样的信念。

信　　念

我们为什么要相信任何东西呢?信念能为我们做什么呢?它们有什么功能呢?我们为什么(想要)具有信念呢?原因在于,世界是无规律地变化的,而生物需要一些适应机制来应对当地环境;这整个工作又不能通过某种永恒不变的结构以及预配的回应(诸如为适应日夜交替之稳定规律的生理节奏)来完成。② 行为的操作性条件反射给予生物以某种适应性,但它存

94

---

① 有些心理学研究表明:个人对消极事件的乐观解释风格——将之归于暂时性的、有限的和外在的因素——比其他的解释模式(将之归因于永恒的、普适的外部因素)能取得更好的个人效果。尽管有这些有利的个人效果——诸如事业成功、幸福以及可能是身体健康——但情况可能是,具有更悲观解释方式的人对这个世界的看法才是更准确的。见 Martin Seligman, *Learned Optimism* (New York: Pocket Books, 1992); 关于准确性问题,参见 pp. 108 - 112,以及 p. 298 页所引的参考文献。这些不那么准确的信念使我们接近这样的人们,他们会相信那种可信值较低——这取决于在其处理机制内的权重以及这些权重的产生方式——但个人效果更好(尽管可能不是有意的)的陈述。

② 对于这些以及相关问题的讨论,参见 Daniel Dennett, *Consciousness Explained* (Boston: Little, Brown, 1991), pp. 173 - 182。

在着两方面的缺陷：一，它不能在新的情境下立即产生新的恰当行为，或者足够迅速地消去旧的、过去曾得到加强但现在已不再恰当的那些行为；二，它在各种环境下无法产生新的但遥远的行为，除非这一新行为经过某种持续的强化链（chain of reinforcement）而与当前行为联系在一起。[①] 信念是可变的，当这些信念是基于理由和推理（亦即基于正反理由的权衡）达至新结论时，那么它们能够调适来符合新的或者正变动的情境，然后对面对这种环境的生物的行为产生有用的影响。

但是为什么有必要去**相信**任何一个陈述或命题呢？为什么要去确定地相信任何陈述，而不只是给每一种陈述赋予各种概率，然后在各种选择情境下依据这些概率而做出最大化效用的行动呢？这就是极端贝叶斯主义（radical Bayesianism）的立场，它也有诱人之处。[②] 感官刺激导致的第一个命题性结论恐怕是（变动的）概率判断，而没有必要构建纯粹的陈述来把这种感官刺激表达为其他陈述的（确定的或很可能的）证据。因为我们不愿依据我们在任何情况下都“相信”的东西行事，亦即孤注一掷，那么，把所有这些判断都只视作是各种信念度（degree of belief），都只视作对那些陈述所赋予的概率，这样不是更准确吗？没有任何信念的概念，人们就没有必要构建对相信什么的接受规则，后者是个艰难的任务。所需要的不过就是不断地修正概率的规则。

不仅如此，极端贝叶斯主义为此付出的代价也并不明显。按照它来说，科学家（或一时期的科学体制）并不接受和相信理

---

① 对此的讨论，参见我的 *Philosophical Explanations*，pp. 703 - 706。

② “极端贝叶斯主义”这个术语是 Richard Jeffrey 的，参见他的 *Probability and the Art of Judgment* (Cambridge: Cambridge Univ. Press, 1992) 论文 1 和 4 - 6。

论或者法则式陈述;相反,卡尔纳普告诉我们,科学赋予这些陈述以特定程度的概率。过去,我们认为科学家相信这些陈述,至少暂时是这样。但这个世纪的科学哲学不断向我们重申:科学理论以及表述的法则最多不过是可能性很高而已。那么,它又如何能够反驳这一点:这就是科学所告诉我们的一切呢?(科学告诉我们某些事物必定是真,但当科学告诉我们这点时,这个陈述至多不过是很可能正确而已,这样说更可行吗?)而且,当你说你绝对确信某事物,而不仅仅是在某种程度上相信它时,极端贝叶斯主义将此翻译为你对它的相信程度为 1:愿意为它下无限大的或者任何有限大的赌注,愿意把你的身家性命押在它身上。(若你不是这样,那么极端贝叶斯主义者将无法理解你的这个主张:相信它而不仅仅是在某种程度上相信它。) 95

尽管有这些显见的长处,但仍很难讲极端贝叶斯主义的立场能够被融贯地表达。各陈述的概率(而不是简单地相信它们)是要适用于选择情境的,但是选择情境就是个人**相信**他所面对的情境,也就是说,他在那里相信自己可以做出各种不同的行动 $A1, \cdots\cdots, An$,相信 $A1$ 可以产生各种可能的结果 Oi(这可能取决于这个世界所处的那种状态 Sj),如此等等。无疑,此人其后会依据这些概率[即 $prob(Oi/A1)$ 或 $prob(Sj/A1)$]而采取行动,但这些发生在有关选择情境的信念结构**之内**。后面的这些能够不只是简单的信念而是适用于陈述的诸概率吗?但个人的概率理论和效用理论的那个设置本身,或是这些理论所要求的背景性补充,就已然包含了信念的存在或把**信念**归因于个人,而他的各种选择就被看作标明了偏好或者概率判断。如果没有个人所处情境的各种信念,那么他的选择就将不能标明他的这些特定的偏好或者概率判断。后面那些概念的理论定义就已经预设了把特定的信念归因于那个人。

这个观点同样也适用于贝叶斯式的基本论证(不管是不是“极端的”),即认为信念度(degree of belief)应当满足概率计算公理,也就是“荷兰赌论证”(dutch book argument):如果信念度表达的是人们愿下的某种赌注,而且信念度不满足概率公理的话,那么这个人就会进入一系列赌博,其中她稳不赚且可能赔。但是,这个论证要起作用的话,这个人就必须相信她所面临的就是如此这般的一种赌博;如果她不相信的话,那么她不会恰当地赌博或行事。因为她愿意下的注乃是取决于她有关该(赌博)情境结构的信念,她在确切的最终结果下会得到何种确切报偿的信念,诸如此类等。如果她仅仅认为自己很有可能面对着这样一个特定的赌博情境,然而她的行为还有某种概率会导致与赌博结构规定的不相同的结果——比如说,当她声明她要把赌注压在陈述 p 上时,天使有 1/100 的可能性会降临人间并改天换地——那么经历很多次这种情境后,就那个(在外部观察者)看来是赌局中唯一相关的陈述而言,她也将不再遵循概率计算公理了。因此,仅当她相信该赌博情境会成立时,“荷兰赌论证”才能得到其结论。

极端贝叶斯主义者可能回复道,在解释或定义信念度这一概念和陈述荷兰赌论证时,对信念的参照是必要的;但尽管如此,信念还是不存在的。信念度确实——能够且应当在解释人类行为中被假定为解释性要素。但是他说,他之所以预设信念,不过是为了引导或者理解我们所假定的那些东西,而一旦过了河,我们就可以拆桥了。 96

极端贝叶斯主义者同样面临着难以克服的实践复杂性。为每一个合式的(well-formed)陈述以及陈述的组合都赋予一个概率,这个工作会多得令人茫然不知所措。但是,若那些与信念不相容的东西都可以不赋予概率——如果自动地被忽略或者被

赋予一个零概率——则信念就可以减轻这个工作。

李维(Isaac Levi)是批判极端贝叶斯主义的,他一直强调信念是用作严肃可能性(serious possibility)的一种标准。一旦我们变得相信某事,那么与之不相容的所有可能性便都可以忽略不计了。信念可以被重新考察;但当人们还持有它们的时候,它们就被认为是确定的,亦即不存在出错的严肃可能性。①李维的理论结构在很多方面让人印象深刻,但对信念的这种处理方式也让他陷入错综复杂的困难。行动者出于某种理由可能不再确信信念 p;但当他还持有这种信念时,他认为 p 是确定无疑的;并且在他还相信 p 且因此认为 p 不存在有出错的严肃可能性时,他随时准备考察那种可让 p 变得不确定的各种理由。李维走了一条迂回曲折且行不通的路径来通过这个雷区,②而且那

---

① 参见 Isaac Levi, *The Enterprise of Knowledge*, pp. 2 - 19 以及各处, and *The Fixation of Belief and Its Undoing* (Cambridge: Cambridge Univ. Press, 1991), pp. 57 - 62 以及各处。

② 你目前持有且因此而相信的一个信念 p,没有出错的严肃可能性,放弃它就打开了这一可能性,即你在日后可能会采纳一项与 p 不相容的信念 q(这个信念 q 甚至有可能就是非 p 本身)。现在你相信那样做无疑是一个错误。既然如此,个人不是应该拒绝放弃当前的信念吗(在他的诸信念尚没有导致矛盾的情境下)? Levi(参见 Isaac Levi, *The Fixation of Belief and Its Undoing*, pp. 160 - 164)对此有如下处理。放弃一个信念("收缩")本身并不会把你导入任何你尚未犯的错误之中,因为并没有加入任何新的信息。然而,放弃一个信念使你处于这样的立场,即在下一阶段("扩展")增加一个错误信念,因为新信念不再会与你那时的信念不相容了。Levi 是这样来解决此困难的:我们在任何时刻都应当只关注下一个行动或下一个决策的后果,而不是关心我们永无止境探询的结果——他把后种关切称作"救世主"的事。可是在下一阶段与时间末端之间还是有相当大的空间的,尤其是下一阶段之后的那个阶段。我们根本就不应该考虑这些阶段,这是完全讲不通的。而 Levi 被迫出此下策,乃是因为他自己的观点认为:相信某事即包括了在所有的语境之下都把它作为具有严肃可能性的标准来对待它和使用它,也就是说,在个人继续持有该信念的时期内都要这样来对待它——而这就使得放弃一种信念极为困难,由此,Levi 被迫走向采纳这种极端短视的权宜之计。

些地雷便是由他自己的理论结构制造出来的;一个外在的观察者很难相信这一旅途是必要的。

我们如何能在极端贝叶斯主义式单纯的信念度和李维理论中所描绘的那种太过顽固且阻碍性的信念之间蜿蜒前行呢?我在此提出一个(非常)尝试性的建议。信念在某一(类)语境下可以排除各种可能性;但在另一种语境下,恰恰不能排除那些可能性。我相信我的一个年轻的新同事不是儿童骚扰者。(要列出这栋哲学楼内非儿童骚扰者的名单,我会毫不犹豫地写上他的名字。)现在语境发生了转换;我要找个人来照看我的小孩两周。那么,此时的一个错误将有很严重的后果——也就是说筹码增大了。现在我就会更加审慎地考虑。这并不是说我此前不相信我同事的清白。在前一种语境之下,对于那些目的而言,我确实是相信他的;我没有考虑过他是一个儿童骚扰者的可能性,或者对之赋予某种概率。但是在这后一种语境之下,因为筹码更大了,我会考虑那事具有何种概率。

极端贝叶斯主义者会欢迎这一点。他说,"恰如我所想,你从不简单地相信它;你总是赋予它以某种概率。"或许情况就是这样的。仅当筹码值得这样做,即当概率乘以相关效用足够大时,我才赋予陈述以某种概率。("足够大"可能不是一个绝对量的问题,也许是一个占总量之份额大小的问题——五元钱对于决定是否在餐馆订一道菜的问题可能产生影响,但对是否买一辆车的问题则不。)我首先要**粗略地**估计该数量是否能足够地大,只有足够大时我才会事实上赋予该事件或者陈述以某种概率(尽管我们一直主张,极端贝叶斯主义者无法对每一个信念都这样说)。

除了相关的效用以外,可能还有其他的语境特征。某些企业,比如说科学和学术出版企业,它们会纳入手稿出版前必须满

足的某些专业标准。这些更为严格的智识规范最终乃是基于最终相关物是什么的一般论证——你无法知晓谁会在何种环境下依据你的出版信息而行动。再或许这些规范乃是帮助界定该企业性质的,有助于我们从它那里获得那类知识。

规则 1 对于日常信念提供了一种充分可信的标准,亦即,与之不相容的任何东西都不能更为可信(虽然这对于信念仍是不够的;至少还要过规则 3 那一关才行)。然而,要在科学内被接受,还要应用一种更为严格的可信度标准:一个陈述不能仅仅是比其他不相容的陈述更可信,还要求达到某种水平的可信度。或者我们应该简单地说,当应用于科学语境时,这个情境将由规则 3 所确定吗?规则 3 本身就是语境性的。科学的标准,一旦为人所知,便倾向于扩展到——某些人也许说"侵入"——其他的语境,首先是涉及社会政策问题的公共语境,其后甚至是人际间问题和个人性问题的语境。

最近开始明朗化的是,人们要主张某种药物的有效性,必须能够得到双盲实验的支持;如果(设定)要改善的病人或判断者知道了是否曾经服药,这会影响其结论。如今在这个领域,双盲实验构成了评价理由和证据的一种标准。(我认为,这个变化是应用"所有相关变量都必须可控"这种恒定标准所要求的。发现了新的相关变量后,就出现了这一改变。)最严格的标准并非在每个领域都是恰当的:我可以合理地做出一个日常的因果陈述,即使我没有进行一个双盲实验来支持它。

98 评价理由的标准到底要有多严格,这要取决于筹码是什么,错误将会有多重大或严重,应用更严格标准的程序需要投入多少精力、时间和资源,还有该事业的一般性质等等。如果存在不同的程序 $P1$, ……, Pn 分别满足不同严格程度的标准,那么我们就可以把选用何种程序的问题本身视作一个决策问题,即计

算每种程序在特定语境下的成本与收益。[1] 但在做出这种计算时,我们又应当使用何种标准或程序呢?我们是否注定逃不掉这个可反驳的循环呢?在何种情形之下应用何种程序是最好的,若要回答这个理论问题,我认为我们想要使用满足最严格标准的最严格程序,从而获得一个一劳永逸的答案,它并不取决于我们当下的那种具体环境的迫切需要(正是它标出了这种理论语境)。

不仅信念要依系于语境,合理性也是如此。把某事称为合理的,这就是做出一种**评价**:其理由是(某类)**好的**理由,且它满足了它**应当**满足的(某类)标准。我们说过,这些标准会随领域、语境和时间等的不同而变化。因此,我们要小心以免做出这样的论断:某人是不理性的,仅仅因为他的理由不能满足我们所能构建的那些最严格的标准。这些理由或许能够满足该语境下恰当的标准,亦即最严格的理论在那里会推荐的那些标准。最严格的理论本身也许认为,要它在那里去满足最严格的标准是不合理的。

如果当某事能够满足它应该满足的相关理由的所有(某类)标准时,它就是完全合理的,那么当某事符合某些而不是所有标准时,或者是在一定的程度上而不是百分百地符合那些标准时,就会存在一种坡度概念(graded notion),这个概念谈及的是合理性程度或某方面的合理性。

在有些语境下,有些东西被认为是理所当然的。这些东西设定了个人要进行活动或做出选择的那个框架,个人在其中努

---

① 参见 John W. Payne, James Bettman, and Eric Johnson, "The Adaptive Decision Maker: Effort and Accuracy in Choice", 载于 *Insights in Decision Making*, ed. Robin Hogarth (Chicago: Univ. of Chicago Press, 1990), pp. 129 – 153。

力想最大化某种功能或让自己的行为展示出某种属性。某个人在语境 C 中把陈述 q 视为理所当然的，就是指当他试图最大化某种功能时，C 中的 q 就是他的立足点。他不会去做一个到达 q 的计算，而是试图从 q 出发到达别处。在语境 C 中我们把陈述 q 视为理所当然的，并且在该语境之中我们得到了信念 r，我们现在就能把信念 r 视为理所当然的，但这只可以是在类似于 C 的语境内，即可以恰当地把 q 视为理所当然的语境之中才行。因为 q 是和 C 绑定在一起的，所以 r 无法自由地脱离为了达致它而在语境 C 之中被视为理所当然的那些东西。①

信念是与语境捆绑在一起的，在此语境下，与信念不相容的那些可能性要么被排除掉，要么被认为不值一提——我们把这

99

① Peirce 认为，在每一种语境之下都有一些不被质疑的东西，亦即一些被视为当然且以之排除掉其他可能性的东西；尽管如此，却不一定有任何特定的东西在每种语境下都看作是理所当然的。Levi 认为，在任何给定的时刻，只要是信念的东西就被看作是理所当然的(尽管某些语境可能会导致你重新考察其中的一部分信念)。我一直认为，这个观点太强了。而笛卡儿的计划就更强了：找到一些信念在所有语境之下都是理所当然的，且永远都不能合理地要求重新考察。有人可能主张，如果 $q2$ 在语境 $C2$ 中被视为理所当然的，且 $q2$ 在该语境下得出 $q1$ 在语境 $C1$ 中是理所当然的，那么认为 $q1$ 在 $C1$ 中是理所当然的就是许可的。如果这种论证要有分量的话，则 $q2$ 必须比 $q1$ 要弱(类似地，$C2$ 也必须更弱或更抽象)。如果向后继续这种过程的话，那么人们也许有望得到这样的一个语境，其中没有任何东西是理所当然的——即笛卡儿的极端怀疑之情境——然而那里有些东西可以得到证成，由此它总是能看作是理所当然的。论者们反对说，笛卡儿在那一极端怀疑情境之中把自己推理的可靠性视为理所当然的了。但我们还可以质疑，他是否将那情境中标出他能得到什么结论的准则视为理所当然的了。看来笛卡儿在此的准则是：如果一个邪恶的魔鬼在 p 为假时不能说服我 p 是真的，那么 p 是可以不受怀疑而接受的。但是“一个魔鬼正在欺骗我”或者“一个魔鬼正在影响我”也能满足这一准则。它仍然肯定不是一个我们在此后应该相信且看作是理所当然的一个陈述。(所以，这个准则在最好的情况下也只是一个信念不容置疑的必要条件。)我们也许可以构建出确立信念合法性的一个更恰当的准则，然而它还是可能面对着诸多反例和困难。笛卡儿看来不仅必须正确且可靠地推理出某一特定的准则得到了满足，还必须正确且可靠地推理出那个特定的准则本身乃是恰当的。

种观点称为“极端语境论”。[1]（是否有一种信念在所有可能的语境之下都被认为是理所当然的，这是开放的。）我们可以使用信念-语境对［bi，Cj］来更充分地确定信念。这种信念-语境对表明，当我（相信我）处于 Cj 语境时，那么我确实是把 bi 视为理所当然的。我是现在这个样子，这个信念就为真，虽然我改变了的话，那么它就不再为真。因此这一信念-语境对标明的只是我现在具有的一种倾向。[2]

只要信念体现着对于行为将会或本会产生的结果的预期，那么信念就会影响行为。行为完成后，这些预期（看来）会被不同程度地证实或者证伪，其后这些结果将修正那些体现着这些预期的信念。有关这个世界的信念向前注入行为，而感知到的

---

① 目标如同信念一样，也会排斥其他选项。那么，给定我所选择的行为是两者的函数的话，那如何来区分信念与目标（它标出了一种结构性偏好或效用）呢？我所选择的行为 B 是我的目标 g 和我的信念——即相信行为 B（很可能）会获得目标 g——的函数。

$$B = f(g, Bel[prob(g/B) = m]),$$

这里 m 值很高。反过来，给定 B，有关 g 的概率的信念又是我其他信念 bel（它们关注达至目标 g 的不同途径）的某个函数 f'。替换后，我们就得到：

$$B = f(g, f'[bel])$$

我们可以假定，函数 f 中也会涉及类似期望效用公式的东西；基于其他的信念（和经验），函数 f' 中将涉及在概率陈述中信念形成的某些公式。因此，信念和目标（bel 和 g）都进入了我们行为的决定，但二者以不同的方式进入，嵌入在不同的函数位置上。［如果我们明确纳入规则 2′中信念的决策理论的（部分）决定，那么上述结论会改变吗？］

② 请注意，极端语境论不同于知识社会学，它并不会削弱自身。假设 RC 是极端语境论学说：所有信念都是在一个语境下持有的。如果某人相信 RC，那么他也是在某一语境中，记之为 Cj，并且在此语境中他视 RC 为理所当然的，并且排除了与之不相容的其他可能性，等等。他不需要认为他在所有语境之下都会相信 RC。说 S 在恰好的 Ci 中是有效的，只要处在 Ci 中，这个陈述本身就是安全的。

这些行为的结果与感知到的其他事实一道,又会或正或负地反馈回信念。贝叶斯论者也接受对于概率的这种反馈,最后导致概率的修正,或许是按照某种条件化(conditionalization)的修正。但是,这种反馈作用也标明了极端语境论与极端贝叶斯主义的区分。极端语境论直接忽略了语境下的某些可能性;这些可能性在极端贝叶斯主义者那里具有某种概率,但是权衡了相关的筹码之后,给定本来所包含的数量,产生的乘积太小从而无法影响决策——因此,他们也(宣称)可证成地忽略它们。然而,在贝叶斯主义的反馈阶段,这些概率看来也必须被修正并要与其他选项保持"同步更新",而这需要在计算上付出大量努力。另一方面,极端语境论者以后也将继续和从前一样忽略这些可能性,至少在他继续处于该语境中时。[①] (在另一语境中,那些可能性和概率也许需要得到考虑。)

哪类信念在何种语境下可以恰当地被认为是理所当然的,对此存在着一般性原则吗?但是这种原则又是在何种语境之下陈述的,那里又可以把什么奉为理所当然的呢?这些原则中是否包含着某种智识理据呢?或者这些原则不过是植入我们祖先的,后又碰巧运作得很成功而得以流传下来的呢?我们估计进化是会植入这样一种语境化信念的机制,还是选择一种足以适合大量不同语境的非语境化信念的机制呢?然而,语境信念理论最终还是发展起来了,存在着它所涵盖的一种信念现象,因而我们就有空间来质疑信念的合理性。我敢说,极端贝叶斯主义理论无法成功地去掉这个话题。

100

---

① 在概率修正时忽略这些可能性会违反某些概率论公理,并因而违反融贯性条件吗?或者语境论者的概率修正只发生在 $1-\epsilon$ 这个区间之内,而 ϵ 尚属未知数(terra incognita)?再或者,正如在信念问题上那样,极端语境论者具有的概率也是附属于语境的,而不是跨语境而不变的?

偏　　见

我们在前面讲过，理性的人不会简单地从她具有的理由净余中外推出有关真理的结论。她会考虑这种可能性，即她知道的这些理由并不是所有理由的代表性样本。因此，理性的人会自察到在她自己的智识功能和所接受的信息中可能具有的偏见。

特沃斯基(Amos Tversky)和卡尼曼(Daniel Kahneman)探讨过这样一种方式，那里人们有时是通过回忆的难易程度来估计某类事例之发生频率的。① 例如，人们料到，在一篇英语文章中随机抽取单词的话，那么以 r 开头的单词会比以 r 作为第三个字母的单词更多，因为人们更容易通过首字母而记起一个单词。但是，人们所想到的单词样本中以 r 作为第三字母的频率在更大的单词样本中却并不是具有代表性的频率。人们回忆信息的方式存在着一种偏见。

心理学家也注意到：在评价或者支持一种信念时，我们并不使用或回忆出我们所具有的相关证据中的一个随机样本或代表性样本。我们最新遇到的证据或者印象最深的证据会得到更大的权重。也有可能的是，当我们获得了能支持某一假说或者信念的一点新信息时，该信息就会从我们的记忆中调出那些符合且支持同一假说的信息。（当前有关视觉不够灵敏的证据会

---

① Amos Tversky and Daniel Kahneman, "Judgment under Uncertainty: Heuristics and Biases", Science 185 (1974): 1124－1131；重刊于 *Judgment Under Uncertainty*, ed. Daniel Kahneman, Paul Slovic, and Amos Tversky (Cambridge: Cambridge Univ. Press, 1982), pp. 3－20，特别是参考他对"可得性启发"(availability heuristic)的讨论，pp. 11－14。

使你回想起自己过去视力不佳的那些情境。)这种信息都契合于一种模式,即一种支持大厦(edifice of support),即使缺失了其中的一根支柱——即便是初始引出其他支柱的那根——余下的支柱也足以支持并维续此种信念或者观点。这种现象可以支持政治生活中的谎言与诽谤的作用。不仅真相永远无法与虚假信息等量齐观(这是我们已经知道的),而且即使在二者等量齐观且排除了那一特定的虚假信息的情况下,也并不能因此就消除已流传的虚假信息的影响,它还是会继续发挥影响。(一个政治候选人可能会指派助手故意去诽谤某一对手,他确信,即便该谎言被揭穿甚至被他所否认,那些谎言仍将起作用。)

有些支持信念的信息被揭露为虚假的之后,那些信念为什么还是不能反弹回原始状态,对此的解释[①]对那些倾向于支持一个假说的*真*信息的效果也有意义。那也同样会唤起记忆中与支持该假说一致的其他信息,但不会同样唤起那些倾向于反对该假说的信息。因此,若缺乏一些专门设计的程序来调出并考虑反面证据,则我们会倾向于高估该假说的可能性和支持力。我们的信念所立足的证据,(一般)并不是我们所能得到的或我们在某种意义上已经拥有的相关证据中的随机样本。某类证据得到一种印象深刻且突出的呈现的话,会在引出其他证据上造成一种偏见,并因此在产生的信念上也造成偏见。[②] 因此,在评

---

① 对“证据贬低后的信念保留”所做研究的讨论,参见 Lee Ross and Craig Anderson “Shortcomings in the Attribution Process: On the Origins and Maintenance of Erroneous Social Assessments”, 载于 *Judgment Under Uncertainty*, ed. Kahneman, Slovic, and Tversky, 特别是 pp. 148–152。

② 心理学者们担忧(为避免实验结果受影响)告诉实验对象的错误信息会持续产生影响,即使他们随后告诉了他们真相,即错误信息被揭露出来了,这些信息还是会有持续的影响。我们当前的反思提出了这种实验的属性问题,那里心理学者一开始给出的就是**真相**(或者是在自然环境下做的实验,不事先给出任何关(转下页)

价一种可能的信念时,不仅仅是要考虑我们所想到的那些支持与反对的证据,还要做出特定且系统的努力来调出我们具有的所有相关证据,无论是支持的还是反对的,这是特别重要的。①

除了我们头脑中想起来的那些信息可能不是我们已有信息的代表性样本之外,我们已有的那些信息也未必是所存在信息的代表性样本。例如,某类信息通过报刊和广播而相对频繁地受到我们的注意,但它们并不是某一主题的全部信息中具有代表性的。新闻资料中所带有的政治偏见或意识形态偏见,使得与此类偏见相左的信息不大可能被传播,而且某些种类的信息甚至更难于被公布或者被刊发,因为编辑认定公众对此种信息不感兴趣或者觉得难以置信。通过我们的信息源而被我们获知的(支持或反对某一特定主题的具体结论的)某类信息偏见,不仅仅体现于该类信息具有的先验概率上,而且体现在该类信息为该来源所提出的条件概率上,给定它成立的话。在我们已有信息的基础上,我们应该得出何种结论,这在部分程度上是取决于:如果所有存在的理由之间的净余是完全不同的话,那么我们(通过传输信息以及理由的信息源)能够获知什么样的不同信息。(贝叶斯框架很适于表达这些问题。)若要考虑人类的信息源所可能带有的偏见,我们也必须考察这些信息源的动机与激励。②

(接上页)于实验的目的)。某人在一个实验中以那种方式回应,且随后跟某个研究员来谈论它,这个单纯的事实可能赋予那个具体信息以一种特殊的突出性(saliency),因此在他塑造日后有关自己的性格与能力的信念中不具有代表性力量。如果真相亦能产生偏见——而不仅仅是虚假才会——那么心理学者就可能有额外的义务,即要抵消他们的实验涉入对人们生活的那种影响。

① 在评估证据的过程中,人们想要纠正的还有其他偏见。一般参见 *Judgment Under Uncertainty*, ed. Kahneman, Slovic, and Tversky。

② 有些物理学家告诉我们,物理学支持宇宙的一种精神观(spiritual view),但是如果这些学者们自身极其向往这种精神性说教,那么这让一个外行人怎(转下页)

102 信念和行动的合理性取决于我们对（我们据以获得理由

（接上页）么去想呢？论者们的报告在多大程度上是得自事实最可行地表明的那些东西，又在多大程度上是他们自己的希冀与愿望的产物呢？如果某物理学家**哀伤地**报告道，尽管他希望情形不是如此，但与其个人的唯物主义前见（preconceptions）相反，他被迫得出，当代物理学得出的教训是，宇宙在根本上是精神性的，那么这给人的印象将是何等深刻啊（但就我所知，这种情况还没有发生过）。

请思考一下《消费者报告》（*Consumer Reports*）这份杂志，它拒不接受任何广告，且起诉任何一家征引对其有利评价的公司，非常小心地防止出现偏见或使人疑其带有偏见的各种可能性。因此，看起来，该报告的读者和订阅者便可以相信，它有动机提供准确而无偏见的信息与评价。但是，为读者服务真的就是它唯一的动机吗？或者说，它是否还有一种动机，即**取悦**这些读者以使他们选择继续订阅这份杂志呢？而假如它只是重申世人皆知的事情，那么这种信息就是没有价值的。因此，《消费者报告》就必须经常报道一些与普遍的想法**相反**的事情，报道那些我们一般认为是最好的产品实际上并非如此，或说最贵的产品实际上还没有另一种产品好等。因此，当读者读到这种故事的时候，她会作何感想：流行的信念常常是错的，因此这份杂志根本没有必要歪曲、诽谤或者掩盖什么事情，由此继续取悦它的读者；还是会想：这个故事是该杂志为了维持其生存而必须时常做的事情之一呢？我倾向于第一种假设，但是给第二种可能性留有余地也是有启发性的。在评价出版物及新闻报道时，我们必须同时考虑到其提供者的激励，捕捉并且表述特定的故事对于其职业生涯会有怎样的助益——比如，报道政治候选人的私生活——还有，这种动机会在新闻的择取以及给报道力度方面造成怎样的偏向，包括这种偏向又会在公共决策的结果层面上如何造成进一步的偏向？

美国的政治竞选中，如今广泛运用焦点访谈的方式（focused interview），精心挑选出群体，以此来找出某个竞选者的何种优点，或者其竞选对手的何种缺点对于这些选民的情感影响甚大。参见 Elizabeth Kolbert，"Test-Marketing a President: How Focus Groups Pervade Campaign Politics"，*New York Times Magazine*，August 30，1992，pp. 18－21，60，68，72。然后，政治口号和政治宣传便会有选择地运用这种信息以确保竞选成功——而获胜者在就职后的实际表现却几乎不受此影响。所强调的论题并不是被发现的重要论题之中随机的或有代表性的样本。如果代表某位竞选人的访谈发现了该竞选人具有一项优点和二十项缺点，而他的竞选对手有二十项优点和一项缺点，那么这位竞选人就会在他的（极煽情且吸引人的）宣传中聚焦于他的那一项优点和对手的那一项缺点。显然，许多选民并不看轻这种不具代表性的宣传，而且那些焦点群体中的成员似乎也并不反对如此选择性利用他们的关注。对于选举的这种重大影响如今已经构成了一个公共问题。

当一本书的潜在购买者看到印在书皮上或者广告中的评语摘引时，他可以认为，该出版商并不是在为我们呈现对于该书的一个随机的或者有代表性的（转下页）

的)这种过程进行判断时的某种自觉性。理性的人将运用某种程序操作且修正她自己的其他程序,亦即修正理由的采样与评价的偏见和要修正据这些理由来做评估的程序中的偏见。①(这个第二层次的修正程序里会不会也存在偏见呢?)大学教育不应当只传授那些获得新思想及有价值的旧思想的技艺,还应当警示学生要提防信息与评估偏见的特定根源,并且告知学生补救或者修正这些偏见的方法。

一般的偏见概念值得我们说一说。我们可以将其区分为两类:第一类是指现有标准的不均衡适用。例如,社会上的歧视就是指对于不同的群体或个人应用不均衡的标准。然而,对

103

(接上页)评语,而是挑选出**最好的**评语,同时注意挑选读者们可能想知道并为读者们所信赖的出处。所以潜在的买家应当纠正在采样中的这样一种偏见且这样想:既然这些是从对该书的各种评价之中挑出的最好的,那么我自己对这本书的看法应该没有那么好。(平装本的出版商们连篇累牍地印刷从各种不同媒体上搜集来的多有重复的溢美之词,他们这样做是想使一个读者相信,这类观点是如此之广泛以至于有理由相信他自己的看法也必定会与此相一致吗?)

谈到了书的问题,我禁不住还想谈一下另外一种现象,这种现象是与我们此处的论题密切相关的。大众传媒通常注重的是那些带有重大经济利益的奖项和奖励,然而它们却非常关注一个只颁发微薄奖金的文学奖:普利策奖(Pulitzer Prize),它是授予小说、历史、传记、诗歌及一般的非小说作品的。对于这种异常关注的一种解释是说,原因在于这一奖项本身的历史和声誉。而我想提出另外一种解释。普利策奖也颁给记者和报纸,所以报纸才会把这些奖励当作重大新闻,值得放在头版头条。如果某人想用很少的奖金为画家或舞蹈编导设立一个奖项的话,那么,我建议他把这个奖也给电视新闻主持人,还有这个节目的制作人也颁一份!

很可能有这样一种联系:不考虑处理故事者的动机,我们就无法根据故事所受处理的重要性来判断该故事的重要性。我个人的看法是,所有文学上的、科学上的、艺术上的以及知识上的奖项都应当受到媒体极大的关注;但是当读者评估这一陈述的时候,我期望他们注意它的来源。

① 心理学文献中也可找到一些有启发性的材料。*Judgment Under Uncertainty*,第 30 - 32 篇论文。另外,可以回忆一下 C. S. Peirce 讨论科学程序的自我纠正性质的著作。

于一种情形的决策或者处理，可以认为有许多不同的标准与之是相关的。是什么决定了何种可能的标准将构成选择的准则呢？当这些得选标准不是在所有可能相关标准中的随机样本时，又是什么决定了它们的权重呢？

为什么选择**这些**标准而不是那些标准？为什么赋予这些权重而不是那些权重？对此的解释如果部分程度上涉及某些人的这种信念，即恰恰是这些标准或者权重起作用排除掉或损害到某些特定的群体，且这点促使他们提出了这些具体的标准，那么在标准的挑选上就是有偏见的。这些标准得到选择就是**为了**排除某些情形的。我们把这种排除称作一种**二阶偏见**(second-level bias)。[①] 如果该标准开始是作为二阶偏见提出的，但可以提出其他的理由支持它，并且该标准之所以得到维持在部分程度上正是由于后面的理由，那么，定义性的和规范性的情境就变得更为复杂了。(我们还可以问，为什么其他的这些准则和目标突然变得显著了，即为什么它们突然就获得了那种权重？在所有可能的准则中间，如果这些不过是**为了**证成那种二阶歧视而提出来的，那么这就构成了一种更高阶的歧视。但我认为，最好还是在二阶层面来考虑这点。)

104

---

① 一些苛刻的大学录取学生时追求“地域分布”，这是1922年哈佛大学所开创的二阶歧视。当时的哈佛校长，A. Lawrence Lowell，公开倡导限制犹太人进入哈佛大学的人数配额制——一个明显地对不同群体应用不同标准的例子。当这种公然的一阶歧视的声明激起了公众骚动时，哈佛大学发现“按地域分布”录入的好处。一目了然，这个目的显然与那些传统目的一样，都是为了限制录取犹太人的，因为当时的犹太人申请者往往都在大城市里面聚居。哈佛的录取程序经历了一种从“一阶歧视”发展到“二阶歧视”的过程。对于这段历史的详细研究，请见 Penny Feldman, “Recruiting an Elite: Admission to Harvard College” (Ph. D. diss., Harvard Univ. 1975)。也可参见 Alan Dershowitz and Laura Hart, “Affirmative Action and the Harvard College Diversity-Discretion Model: Paradigm or Pretext?” *Cardozo Law Review* 1 (1979): 379–424。

极为老到的观察者有时不提及有可能存在的附属的二阶歧视,而直接高明地讨论一阶歧视的问题。统计学者经常引用的一项研究,即考察(位于伯克利的)加州大学在某一段具体时期内,研究生院在录用过程中是否歧视了女性申请者。① 依评价标准来看(本科成绩、专业课程数目、研究生考试得分,等等)女性申请者看来与男性申请者同样合格,然而,该研究院对女性申请者的录取比例却远远小于男性申请者。这必定是一种歧视情形吗?(统计学者说)不是的,因为如果我们一个系一个系地分别考察录取情况的话,我们就会发现在每一个系内男女申请者的录取的百分比大致相等。没有哪个系有歧视。那么男女申请者录取的总体百分比又怎么会如此不同呢?原因在于男女申请者并不集中于相同的系。有些系比另外一些系的录取比例低,而女性申请者更集中于这些系上。(简单来说:假如大部分女性申请的是只有10%录取率的系,而大部分男性申请的是录取率为50%的系,那么整个研究生院录取的男女生比例就是不同的,即使每个系的记录表明它在男女的录取的百分比是相同的。)(统计学者)案子结了:不存在歧视。

好个干脆的论断。但是这项研究所能告诉我们的只是不存在一阶的歧视;证据上显示不出歧视来。我们仍旧可以问,为什么不同的系要有不同的录取比率呢?很有可能这是因为各个系的可录取名额与其申请者数额之比例是各不相同的,而女性们碰巧申请了可录取名额与申请者数额比例相对较小的那些系而已。但是,这果真是"碰巧"发生的吗?为什么研究生院系的规模与合格的申请人数不成比例呢?是什么决定了一个系的规模大小,即什么决定了学校投入多少资源用于资助每一个系中的

① P. Bickel, Eugene Hummel, and J. W. O'Connell, "Is There a Sex Bias in Graduate Admissions", *Science* 187 (1975): 398 - 404.

教师岗位、研究生奖学金等等的开支呢？原因也许有很多。然而，假定一些研究生系受到的资助不足乃是因为这些系有更多的女性研究生。假定因为这个原因，在分派研究项目资金时，学校领导或更广泛的社会公众会认为这些研究课题不太重要。或105 者假设其他的研究项目获得了更多的资助，有能力录取更高的比例，因为这些项目属于“男性”项目，因此被视为有更重要的社会意义。有些系本来应该更大，要不是**因为**它们有更高比例的女性申请者的话；或者说其他系的规模能够保持更大，是因为男性申请者的比例更大。我并不是在宣称事实就是如此，只是在描述一种并非完全讲不通的可能性。在这种可能性下，伯克利的统计数字就将符合一种二阶歧视的模式。单独这些统计不会证实这点；要发现它就必须对一个有关学校组织的结构性问题进行调查，亦即，为什么不同的系录取学生的比率会不同？[①]

在二阶歧视的情境中，应用的标准（或赋予它们的权重）并不是所有相关的可能标准中的随机样本。不仅如此，任何用来证成这些具体标准（或其权重）的准则也并不是所有可能相关的评价准则中的随机样本（或得到了客观证成的一个子集）。这里

---

① 另一种非歧视（或相对不那么严重的歧视）的草率结论是基于由 Thomas Sowell 所做的统计，他的论证如下。几乎所有的白人都不能讲出诸多黑人亚族群（subgroups）之间的差别，甚至都不会去注意这些差别，因此可以想见白人也同样不会在他们之间有所歧视。但是，加勒比群岛上的美国黑人和那里的美国白人就挣了相同的平均工资。因此，难道不是其他亚族群自身基于文化的特性，而不是白人的歧视，使得他们所挣到的平均收入低于白人吗？参见 Thomas Sowell, *Civil Rights: Rhetoric or Reality?* (New York: William Morrow, 1984), pp. 77 – 79。

然而，有些白人亚族群的平均收入高于一般的白人收入水平——比方说，斯堪的纳维亚裔白人和犹太人。或许如果不是因为存在歧视的话，加勒比群岛上的黑人本可以比白人挣得更高的平均收入。因此，有可能存在着对所有黑人的歧视，使得他们的收入低于本来会有的收入。存在一个黑人亚族群处在白人的平均水平，这一事实并不能够排除对全体黑人的重大歧视。

的取样是故意有偏见的。在另外一些例子中,问题就不是这般显而易见了。我们想一下这个争论,即文学理论者有关文学经典的性质以及纳入经典要符合什么样的条件。对于有些女性作家、少数派作家以及来自其他文化的作家们,即使他们如今已达到了一直以来使得主流的男性作家得以登入文学经典殿堂的那些标准,但却仍然被排斥在外的话,那么就可能存在着一种针对她们的一阶歧视。然而,即便现有的那些标准公平地应用于所有作家,但我们依然有可能在经典的形成中发现有二阶歧视。为什么恰恰应用**这些**标准?是否还有其他的优点或者准则,使得一部作品值得我们用完全相同的方法加以认真研究?是否还能设计出其他有趣且卓有成效的方式来研究及阐释这些不同的作品呢?这并不意味着一个人必须在评价过程的一开始就说出那些标准到底是什么。在亚里士多德写出《诗学》(*Poetics*)明确地阐明埃斯库罗斯(Aeschylus)与索福克勒斯(Sophocles)的戏剧所满足的标准之前,希腊人早已知道此二公的戏剧是不朽的传世佳作了。[1] 探寻新的标准未必要否定既有标准所赞赏的那

---

① 对这种二阶任意性(second-level arbitrariness)——它不一定是歧视——的矫正,应与另一个不时被提出的目标区分开来,即通过把美国那些受压迫(或自认为受到了压迫)的某些少数派(minorities)的作品纳入教材,这样以提高他们的自我形象或者他们的外在形象。请注意,把这些群体当中的一些作家们纳入教材,这样做还是能够存在**教育**目的的,即便这些人的艺术成就没有那些纳入的极伟大的作家那么好。但他们仍然要比要读这些教材的绝大多数学生都更加敏锐,且天赋更高。有些学生不承认有些女性和少数群体中的人比他们自己更聪明、更敏锐及更有天赋,让他们记住这一点,这本身是具有非常重要的教育功能的。

在 *Philosophical Explanations* 一书的最后一章中,我描述了一种价值和意义不断交替的过程:先确立起统一性,再向外接触且涵盖那种会打破统一性的更进一步的、多样性的新材料,然后再确立起新的、更加宽泛的统一性,接着向外接触到更为广泛的材料,如此类推。那么,相信暂时(受威胁的)统一性的人们将会卓有成效地把"多元文化主义"看作这一进程当中的一个部分,而不是其终结阶段。

些优点。

我认为,正是在诸多可能的准则和标准之间进行选择这个领域内,知识社会学颇有用武之地。对任何一种信念都存在着诸多可能的正反理由——与有争议的社会性或规范性问题相关的每种信念则肯定如此——并且存在着对这些理由进行评价的诸多标准。没有人能毫无偏见地找出所有可能的理由,并赋予它们同等的权重——包括我们中间会努力地考察那些抵消理由的那些人——而且任何人所接受的理由都不是所有可能的理由之中的随机抽样。看来可以合理地认为,[在卡尔·曼海姆(Karl Mannheim)传统内的]经典知识社会学者所研究的要素——诸如阶级地位、教育水平、群体纽带网络等等——将影响到具体的个人在各种可能的合理理由中会注重其中的哪一些,亦即哪些理由与标准对他而言是凸出且被看重的。(最近,其他论者将关注点放到了性别、种族和性取向上。)当我们面对一种复杂的社会情境时,处于不同社会地位的人会注意且聚焦于不同的侧面,并由此援用(适合这些方面的)不同原则进而达至不同的结论。每种结论都可能是正确的:每个人注意到的都是能够支持各自(对立的)结论的理由。得出每一种结论的人也都可能是理性的:有理由地相信,基于证据得出结论,援用能够得到很多支持的信念与评价准则。但每个人都只是依据某一部分理由来相信某事,都只是在部分证据的基础之上得出结论的,都只是援用部分得到很多支持的准则。(社会学家所研究的)社会因素缩小了本要得到考虑的那些可能的相关考虑的范围。在这个缩小的范围之内,我们的信念与行为可以是合理的。即便不是相对于这个范围,那也不能说所得出的信念和行为就是完全不合理的——既然有某种理由在支持着它们,它们至少是初定地(prima facie)合理的。

106

在序言中，我们曾瞄了一眼合理性本身是有偏见的这一主张。反过来也许可以这样主张，虽然合理性旨在获得其所追求的目的这一点上并没有可反驳的偏见——它尽其所能来矫正这种偏见——但它在追求的目标本身以及其追求的方式上，是有偏见的。合理性不是排斥感情、激情以及自发性吗？而这些东西不又是人们生活中有价值的成分吗？然而，合理性是可以追求这些东西的。即便是在决策理论中的合理性也能推荐不假思索或者不经计算就做出很多决策，只要这样做的价值大于其不那么反思的决策将会带来的损失即可，或者只要这种计算过程本身将会在本质上扰乱其他有价值之关系（诸如爱与信任）的话。[①] 确实，如果合理性本来这样推荐的话，那这将是基于它所认定和评价为好的理由来做的，这样那种推荐便不是轻率的了。然而，这与那种计算推荐并不是一样的。对理由的回应并不要求要明确地考虑它们。理性可以是谦抑的，有时会选择避到一边，甚至在某类环境下几乎总是如此。

---

① 参见 Robert Nozick, *The Examined Life* (New York: Simon and Schuster 1989), pp. 76 - 83, and Olver Williamson, "Calculativeness, Trust and Economic Organization" (March 1992 年刊出了 1992 年 4 月在芝加哥大学法学院的法与经济学论坛上的一个讲话)。一个人不对朋友的可信度进行精确计算，就是相信他，但这一点并没有蕴含着充分强的反证据，使得不可能让他相信他的朋友是不值得信任的。

4. 进化的理由

信念或行为的合理性既是回应其正反理由的问题，也是据之而产生这些理由的过程问题。为什么合理性涉及理由呢？这里有一个答案：信念和行为要具有某些属性(properties)，比如说真理或满足欲望等；如果它们能回应所有正反理由，那么它们具有那些属性的可能性就更大。(有些其他的过程根本就不考虑或权衡理由，它们甚至有可能更可靠地达到那个目标吗?)在达致我们的认知目标上，暂且不论考虑理由是不是最有效或者最可靠的方法，为什么它能是有效的呢？什么因素把正反理由与这些目标关联在一起呢？什么因素使得某事成为一个理由呢？在存在的混杂不清的信息当中，又是什么构成了作为信念或行为正反理由的事物呢？

经过达致、维持和修正信念的一般性过程，我们逐渐持有了信念。对于不同类型的信念，甚或只是不同场合下的同类信念，我们采用的过程都可能是不同的。在特定的场合下遵循特定的过程，我们可能会由此得到一个真信念；运用该过程将(始终或概率性地)引起我们相信该真理。理由的运用能够起作用，让过程成为我们相信真理的概率性原因，且各种推理程序之间的差别也能影响过程的效率或有效性。只有当 r 的真与 h 的真之间存在着某种关联的时候，出于理由 r 相信 h 才会有助于我们获得真信念这一认知目标，这看来是讲得通的。理由与理由所支

持事物的真理之间存在着这种关联,正是这种关联解释了出于理由而相信与相信的是真理之间所存在的那种关联。那么,理由与理由所支持事物之间的关联又有什么样的本质呢?

理 由 与 事 实

就信念的各种理由而言,哲学著作中包含两种观点。第一种观点是先验观,它认为:假说 h 的一个理由 r 与 h 之间具有 108 某种关系 R,理性官能(faculty of reason)能够领会到(apprehend),这种(结构性的?)关系构成了一种支持关系。理由就是心灵(mind)有能力辨识的东西。① 问题在于:为什么当 r 为真且 r 与 h 之间处于某种关系 R 时,我们就预计 h 实际上也为真呢?如果回应只是说 r 是支持 h 的一个理由,那么我们还可以继续问,当 r 为真时为什么 h 实际上也(常常)为真,我们对此能给出何种解释呢?②

第二种观点是事实观,它认为当 r 与 h 处于某种偶然的事实性关系时,r 就是 h 的证据。在《哲学解释》一书中,我主张证据性关系是一种事实性关系,并且依据证据与假说之间的那种追踪关系并依据那种关系的概率式接近(probabilistic approximations),提出了这种事实观的一种论说。③ 尽管在此我并不想坚持这种

---

① 然而,即便我们的心灵无法认识到这些关系之间的联结(concatenations)或扩展,这些关系还能构成理由吗?也许这种理性关系是递归可数的(recursive enumerable)但不是递归的吗?

② 参见 Nelson Goodman, *Fact, Fiction and Forecast* (Cambridge, Mass.: Harvard Univ. Press, 1955), pp. 65 - 66,他对先验观也提出了一个与此类似的问题。

③ *Philosophical Explanations* (Cambridge, Mass: Harvard Univ. Press, 1981), pp. 248 - 253. 其他论者也提出了支持度是偶然性的各种观点。一(转下页)

事实性关系的那种特定论说。

当这种事实性关系得到恰当的规定时，与 r 一道，那么"知道这种关系成立"就构成了相信 h 的一个理由。但是，在不知道这种关系成立的情况下——尽管它确实成立——并且 r 与 h 之间也没有任何自明的结构性关系，那么 r 还构成相信 h 的一个理由吗？事实观似乎遗漏了那种最能打动先验观的因素，亦即在特定情形中，这种理由关联看起来（近乎）是自明的。（要注意，**证据**这种概念可以契合于一种纯粹的事实论说，即使**理由**概念不行。）

我建议我们组合这两种观点。h 的一个理由 r 是一种与 h 具有某种（这一点还有待进一步的理论来确定）事实性关联的东西，尽管若 r 与 h 的内容处于某种结构性关系，看来更能打动我们，让我们认为：给定 r，h（更）可信。除了经验外，这种理由关系看来是一种支持性的事实性关联。（对于**支持**这个术语，你可以把它替换为理性能力的倡导者所喜欢的任一描述。）

对于我们按照一种事实性关联行动而言，可能存在着不同的基础：（假如这种事实性关联在过去是成立的，且对那种事实-行动系列存在一种进化选择的话）该行为可能是被预配的（prewired），或者是操作性条件反射的。但是还有第三种基础：依凭**理由**行事包含**认识到**内容之间有种结构性关系的关联。这种认识本身可能一直是有用的，从而得到了选择。某种事实性

（接上页）定数量的"P 是 Q"的例子能在多大程度上支持"所有 P 都是 Q"这样一个假说，这将取决于我们现在相信有多少种 P 与 Q 相关，也就是说，取决于我们 P 所是的那类事物的（相关）变体范围与 Q 所是的那类事物的（相关的）变体范围有多大。见 John Holland, Keith Holyoak, Richard Nisbett, and Pual Thagard, *Induction: Processes of Inference, Learning and Discovery* (Cambridge, Mass.: M. I. T. Press, 1986), pp. 232 - 233, Norman Campbell 所预见的一种观点，*What is Science?* (1921; rpt, New York: Diver, 1952), pp. 63 - 64。

关联在我们看来是自明的证据，这种特性得到选择和偏爱，是因为在一般情况下，依据这种（确实成立的）事实性关联而行事会增强适合度。我不是表示，得到选择的正是这样的能力，它能认识独立存在且有效的那种合理关联。而毋宁说，存在一种事实性关联，并且过去在生物之间存在一种选择，使得那类关联看来是有效的，即注意到那类关联，且这种注意导致某种额外的信念、推理等等。存在一种选择让我们把某种是事实性的关联承认为有效的，也就是说，使得它们在我们看来不仅仅是事实性的。①

109

如果某类样本足够频繁地类似于其总体，那么从样本到总体的概括，或从该样本到下一个出现的成员的概括，就通常是对的；如果这样的推论对某些存在物看来是显而易见和不证自明的话，那么他们也会频频地得出那些真理。这个例子涉及归纳推理的一般过程。[科斯米迪（Leda Cosmides）和图比（John Tooby）曾研究过这样一种可能性，即对我们在进化史上频频遭遇的特定类型的情境而言，有些专门的推理机制是有效的，由此得到选择。②] 要注意这种进化选择有可能是鲍德文效应

---

① 鉴于近来有关适应主义（adaptationism）的争论，如果这种假说在对大脑特征的进化选择上不要求过多特殊性的话，那么它或许是可欲的。参见 Stephen Jay Gould and Richard Lewontin, "The Standards of San Marcos and the Panglossian Paradigm: A Critique of the Adaptationist Programme", *Proceedings of the Royal Society of London*, B 205 (1979): 581 - 598; 也参见各种争论最优性的论文，载于 *The Latest on the Best: Essays on Evolution and Optimality*, ed. John Dupre (Cambridge, Mass.: M.I.T. Press, 1987), chs. 4 - 9。

② Leda Cosmides and John Tooby, "Are Humans Good Intuitive Statisticians After All?" (即出); Leda Cosmides, "The Logic of Social Exchange: Has Natural Selection Shaped How Humans Reason?" *Cognition* 31 (1989): 187 - 276; John Tooby and Leda Cosmides, "The Psychological Foundations of Culture", 载于 *The Adapted Mind*, ed, J. Bardow, L. Cosmides, and J. Tooby (New York: Oxford: Oxford Univ. Press, 即出), pp. 19 - 136。

(Baldwin effect)的一个例示(instance)。① 在这种具体情形中，如果某些人的“配置(wiring)”使得一种关联看来更接近于自明的话，那么他们就能够学得更快，由此获得一种选择性优势，然后他们留下的后代也会对这种关联有差不多的自明度。随后，一代又一代，这种关联就会越来越具有自明性。

请注意，这个观点(像前面一样)也给我们提出了归纳问题：尽管某种事实性关联在过去成立，并且选择使得成为我们这样的生物，把这种关联看作推理的有效基础，但这种事实性关联在今天以及将来还会继续成立吗？这种关联将继续成立，这一陈述本身或许就与我们已知的其他事实看来处于一种有效关联中，但那个进一步的(或许是我们以之开始的那种)关联继续成立吗？它在我们看来是自明的，这一点并不能保证它将继续成立，因为这个“看来”只不过是由于它在过去一直有效而产生的。

不仅如此，基于此种论说，某事在我们看来自明地为真，这并不能保证它在严格的意义上曾经为真过。通过类比，我们可以考虑现在对欧几里得几何学的看法：几乎对于所有实践性目的而言，它都是足够正确的；它与小的常曲率(constant curvature)空间之间有着微小的差别，尽管无法察觉；但对于我们所说的“物理空间”来说，它在严格意义上并不为真。如果当初欧几里得几何学确实是因为看来自明地为真而得到进化选择的，那它就已经很好地帮助了我们的祖先。要是选择的是其他几何学，那他们就会一无所获。(在该环境下)相信其他的几何学并自动地据之做出推论，这并不会赋予先辈们任何选择优势。这是因为，依据所要投入的神经学资源而言，这种替代的几何学

① 对于鲍德文效应的一个讨论，请见 Daniel Dennett, *Consciousness Explained* (Boston: Little, Brown, 1991), pp. 184 - 187。

知识要看起来是自明的,耗费的成本会太过高昂。并且"直觉
110 到"这种替代几何知识或许超出了当时(在没有逐步的进化选择压力的情况下)人们从既有的基因禀赋的随机变异(random mutation)中所产生的能力范围。那么这种几何学,尽管在严格意义上为真,但不会被选择成为在我们看来是自明的。由于欧氏几何学是相当接近真理的,且让它对我们看来是自明的还具有诸多附属优势——包括推理快捷、相信有用的(近似)真理和避免其他多种错误——我们能够想象,欧式几何的"看来自明性"是被选择的;我们还能够想象,选择那种(正确的)几何学作为我们的感知形式。尽管如此,严格说来,作为一种物理空间理论,我们现在认为欧氏几何学乃是错误的。[而普特南(Hilary Putnam)问道,过去的几何学还能是别的样子吗?]我并没有主张,这个进化故事**就是**为什么"欧氏几何学在人们看来是自明的"真论说。它只是通过类比来提出这个观点——在一般情况下是我们已经知道的一个观点,但在**理由**、**推理**以及**证据性支持**这一领域中却往往被忘记掉了——一种关联成立的表面自明性(鉴于其他某种明显的结构性特征或关系)并不确保它事实上确实成立。

我们要对演绎规则和逻辑原则本身的"自明性"提出类似的主张吗?它们是必然的吗?还是传统上所有先验知识都会被扫到那个"进化仓"之中去呢?有些论者宣称,逻辑原则既不是必然的,也并非是先验可知的;即便我们目前还没有建构出替代物,但某些难处理的现象——量子论之类的东西——可能促使我们去修正甚至是我们的逻辑原则本身。①我的观点更为温和。

---

① 参见 Hilary Putnam, "Three-Valued Logic", and "The Logic of Quantum Mechanics", 重刊于他的 *Philosophical Papers*, vol. 1: *Mathematics*, (转下页)

若要解释为什么这种原则在我们看来是自明的,我们不一定要援引其必然性。只要它们是真的,即便只是偶然地为真,甚或仅仅是"足够真"(回想欧氏几何学的例子),并在足够长的历史时期内一直被人们视为真的,从而在我们的进化禀赋上留下了痕迹,[①]这就够了。这个立场并不受制于奎因的这一个强有力反驳,即所有的逻辑真理不能将其真理性归于约定,因为我们需要援引逻辑原则来派生出约定的无限后果。[②] 我们认为逻辑原则确实为真——无论如何是足够真的,且或许就我们所知,是偶然地为真的——进化过程植入的(不是逻辑原则的真理,而)是它们的看来自明性。由此在推出它们是作为植入为自明的后果

111

(接上页)*Matter and Method* (Cambridge: Cambridge Univ. Press, 1975), pp. 166 - 97, and W. V. Quine, *Philosophy of Logic* (Englewood Cliffs, N. J.: Prentice-Hall, 1970), pp. 85 - 86, 100。如果正如奎因所认为的那样,逻辑学与数学始终是经验科学的延续部分,是经验科学自身的一部分,那么对于这些逻辑法则或者数学理论为什么成立,我们为什么不去寻找更深层的科学解释呢?物理学家不断地为当前所知的最深刻法则找寻更深一层的解释,然而在逻辑学家那里找不到任何可与此相媲美的探究活动。只是因为逻辑学家们已经发现了最根本的法则,这是说不通的。为什么这种终结在逻辑学史上这么早就发生了,而在物理学中却至今尚未发生呢?奎因在谈话中指出,他在 *The Root of Reference* (LaSalle, III.: Open Court, 1973, pp. 76 - 78)一书中主张,有些逻辑法则是分析性的;当我们知道那些构成词的意义时,就知道它们为真了。(说得不错,奎因!)但这看起来并不足以解释,为什么在逻辑上没有继续对支持命题计算(propositional calculus)或者量化理论(quantification theory)等的各种真理(truths)之真理性的解释性探究:把这些各自保真的(truth-preserving)短小演算步骤在一个长长的推理链条中反复叠加(iteration)也是保真的,这不是一种数学(亦即非分析性的)事实吗?

① 但这个讨论也许会继续:它们为什么是真的,为什么在足够长的历史时期内被人们看作是真的,从而对我们的自明性感觉产生此种进化上的影响,什么因素能解释这一点呢?若偶然观(contingent view)不能提出任何可行的更深层的解释性假说,那么必然性的倡导者也就不能简单地宣称:它们如此长久的成立是因为它们是必然的,而对它们为什么是必然的则不知所措。

② W. V. Quine, "Truth by Convention" (1936), 重刊于 W. V. Quine, *The Ways of Paradox* (Cambridge, Mass.: Harvard Univ. Press, 1976), pp. 77 - 106。

时，可以毫无障碍地假定它们为真。如果某些陈述是这种类型的，即一旦它们是偶然性的事实的话，就会导致对其自明性的强烈直觉的进化选择，那么对于这些类型的陈述而言，我们在直觉上具有的那种强度和深度并不能为它们的必然性提供强有力的证明。①

哲学家一直以来面临着这样的任务，即要为理性(Reason)提供根据，为我们认为自明的东西提供根据。休谟的归纳问题就是要找到**合理的**论证以得出这个结论，即理性，或体现于归纳推理中的那部分理性，(很可能)是起作用的。即便这个归纳问题能被解决，②仍然存在着这个问题：**那种**合理论证的根据是什么，即为什么我们应当信任**任何的**合理论证。这也曾经是笛卡儿面临的问题——为什么受自然理性之光普照的那些自明命题，必定会符合实在(correspond to reality)呢？——并且这还引出了大量文献来讨论"笛卡儿循环"(Cartesian Circle)。③ (有

① 这两种区分是相并列的。第一种区分是两种证据的区分，作为一种现实关联的证据(evidence)和作为一种(近乎)自明的先验关联的证据。第二种区分是两种合理性的区分，一种合理性是作为从一种在现实可靠的程序当中产生出来的，另一种合理性是由相互重叠和相互关联的各种陈述、观点与推论所组成的某种严密的结构中产生出的。在上述两种情况下，我们都有一个区别于合理方面的事实方面；在这两种情况下我们都想让两个方面是相配的，并且当合理方面不是铆定于一个现实关联时，我们都是不安的。当我们相信那个合理的方面是附着于事实方面时，则我们就会心安理得地认为一种合理例示(instantiation)本身是有价值的。而如果二者看来是割裂的，亦即当一种合理性模式似乎不再能反映现实或是达致现实的一种方法时——正如学术争论传统中所发生的那样——则该模式也就失去了吸引力，它看上去不再那么优美，也不再具有内在价值。

做个对比：当所要求的或正当的行动也有相当好的后果时，那么我们对伦理学中的义务论立场心安理得就容易得多了。

② 如果归纳推理是合理的，那么就会存在一种合理的论证，亦即归纳论证；这会由于循环被拒绝。因此，这个问题被认为是要用**其他**部分的理性以一种非循环的方式论证这一部分理性，即归纳推理。

③ 另请见第155页注释①。

趣的是我们要注意到，笛卡儿最终是把他对于合理论证的信任立基于对另一种存在——即上帝——的信任之上。）康德认为唯理主义者（rationalists）不可能表明，为什么我们的知识或直觉（在我的意义上，这是指我们的“理性”）符合对象（objects），并且他还表示——这就是他的“哥白尼革命”——对象必定符合我们的知识，亦即符合我们的直觉能力的构成。① （因此，我们的知识并不是物自体的知识，而只不过是关于经验实体的知识，因为经验实体乃是由我们的构成所塑造出来的。）

康德说，如果理性与事实是独立的因素，那么两者为什么是彼此相符合的，唯理主义者对此就无法提出任何有说服力的理由。为什么这两个独立的变量是有关联的呢？康德于是提出，（经验）事实并不是一个独立的变量；事实对于理性的依赖解释了二者之间的关联性与相符性。但是还存在第三种选项，即**理性**才是因变量，它是被事实所塑造的，并且理性对于事实的依赖解释了它们之间的关联性和相符性。我们的进化假说提出的恰恰就是这样一种选项。理性能告诉我们有关实体的东西，这是因为实在塑造了理性，选择了那种看来是“显而易见的”东西。 112

我们已经说过，这种观点只能解释过去的关联；它无法保证将来的事实会契合当前的理性。而且，进化解释本身也是（在部分程度上）我们运用理性支持一般的进化论，支持这里对它的具体应用而得出的。因此，这无法为理性提供一种独立于理性的证成。而且，即便这种观点把独立于理性的事实用作理性的根据，这种根据也无法独立于理性而为我们所接受。因此，这种论

① Kant, *Critique of Pure Reason*, trans. Norman Kemp Smith (London: Macmillan, 1933), 第二版的序言。

说并不属于第一哲学；它是属于我们目前不断发展的科学观。①它也不是用来满足康德的这个准则的，即“以任何方式类似于假设的东西都是违禁品”。②

我已经说过，我们最好把哲学视为对理性的爱——这不是对智慧的爱，而是对推理的爱。即便是那些希腊怀疑论者和英国经验论者，亦专注于推理并以推理为荣，尽管他们的推理往往会减小和削弱理性及推理过程自身的权威——因此他们要么是回避要么是面对实用主义悖论。哲学家要给理性以根据是为了尽力保护他对理性的这种爱。（或是为了保证这个爱忠于他？）但接受这种对理性力量与表面魅力的进化解释会减弱这份爱吗？未必如此。当我们知道感觉器官有一种进化解释时，难道我们的耳朵与眼睛的价值就降低了吗？尽管如此，有些哲学家还是把理性的声音视作对与偶然性相对立的必然性的宣告，认为它提供了一种现实性(actuality)所无法涵盖的那种（理解）门路(access)；他们也许会觉得他们的爱所带来的那种特殊慰藉被剥夺了。

这种进化论论说表明了为什么某事表面上看来是显而易见的。网络模式内的各种成分向前输送时具有可变权重，且要受制于纠错规则，理由在这种模式内构成了一个更为宽泛的范畴；注意到的事实关联在这里是反映在变化的连接(linkage)和权重之中。因此，我们并没有局限在仅仅是进化过程所灌输的理由之中。这是幸运的。进化也许只把一些近似真实的东西植入为显而易见的，但对于我们随后的某些目标而言，这也许是不够 113

① 比较 W. V. Quine, *Word and Object* (Cambridge, Mass.: M. I. T. Press, 1960) ch. 2。

② Kant, *Critique of Pure Reason*, 第一版序言。

的。同时,鉴于我们没有完美的准确性——无论如何,以一个合情理的成本是得不到这种准确性的——那么进化或许更易于犯这些而不是那些错误。有两种类型的错误,一种是错误地相信有一只老虎出现,从而没必要地跑开,另一种是错误地相信没有老虎出现而一动不动,后种错误对适合度会有更大的危害效果。因此,进化就可能选择那些更易于犯前一种错误的机制。[①] 在不同的环境下,避免某个特定的错误可能是不那么重要的。如果一个适应性的处理系统所具有的初始权重是允许被修正的,那么它就有能力获得更高的准确性。

回想一下我们先前所做的区分:(1) p 是要相信的合理的事情,(2) 相信 p 是要做的合理的事情。自然选择对后者起作用,并且只在二者之间存在关联时,它才为我们选择一种也适应前者的认知机制。确切地说,自然选择首先作用于:(3) 行动 A 是要做的最能增强适合度的事情;或者再确切点说,A 产生于这样的能力,运用它们一般能够增强全面的适合度。因为你的所为是你的所信以及你的动机与效用的产物,所以行为倾向如何实现是有变动余地的。因此,我们不是通过信念机制,例如把含糊的数据马上看作表明出现了危险事物,而是通过动机机制,例如厌恶感导致人们避开蛇,得以避免某些危险。

增强全面适合度的要求产生了对于近似真理而不是严格真

---

① 参见 Robert Nozick, "Experience, Theory and Language", 载于 *The Philosophy of W.V. Quine*, ed. Lewis Hahn (Lasalle, III.: Open Court, 1986), pp. 340 - 341, and Stephen Stich, *The Fragmentation of Reason* (Cambridge, Mass.: M. I. T. Press, 1990), pp. 60 - 63, Stich 继续论证,因为犯错具有不同的代价,当机制的其他优点压倒了不那么可靠这个缺点时,进化过程所选择的那些认知机制就不能最可靠地揭示真理。

理的选择。知道这一点后，我们就可以*更明确*(sharpen)我们的目标和其程序。如果要为了行动中的信念有用性而选择，且真理是一种一般都会支持这种有用性的属性，那么我们就可以集中于真理而不仅仅是有用性来构建各种程序。而且，我们还可以明确真理这一概念。或许并非*所有*的有用性都标出真理或为它所支持——进化的理论化本身就可以告诉我们，有不同种类的东西可以支持有用性——而且我们可以说，真理乃是支持有用性的一个子类的因素。一旦我们自觉地认识到了这一点，则我们将能够改善我们既定程序的准确性。

我们再来考虑这两者之间的关联，一个是信念形成程序产生真理的那种可靠性，另一个是以理由为基础而对信念的那种获取。如果进化要选择可靠的信念形成机制，如果出于理由而相信是这类可靠机制的成分，那么由此导致的生物关注且着重的可能是理由而不是可靠性。这一着重是指导它们得到可靠性的方式，而可靠性本身却不是它们的着重点。类似地，曾经有对114 这样一种心理机制的选择，这种心理机制不是选择对全面适合度本身的直接关注，而是在统计上与最大化全面适合度相关联。在可靠性与理由发生实际冲突且人们意识到了这一点的情况下，人们会更倾向于理由而不是可靠性；在可靠性独自出现的情况下，它看来只是一个不充分目标。但选择所要服务的却恰恰是可靠性。对于理由的关注，在过去是因为与得到真理的可靠途径的关联而出现的，但如今却可自由变动了。

适合度与功能

我们要更仔细地看看进化解释的结构与轮廓。各种文献告诉我们，进化涉及适合度的可遗传变异，即把生物之间不同在非

随机差异生殖上发挥作用的母体特征传递给后代。① 适合度并

---

① Richard Lewontin 告诉我们，成熟的群体遗传学理论有如下结构(Lewontin, *The Genetic Basis of Evolutionary Change* [New York: Columbia Univ. Press, 1974], ch. 1, esp. pp. 12 - 15, 这段剩下部分的描述皆来自此书)。它包括族群在 $t1$ 和 $t2$ 时刻的基因描述(genotypic description)$G1$ 与 $G2$,以及从一类变为另一类的转变规则：一组给出表型分布的超基因遗传(epigenetic)规则*,这些表型是由不同的基因组在不同的环境下发展而形成的;种群在一代跨度内的交配规则、迁徙规则及改变表现型排列的自然选择规则;允许基因组分布与任何表型分布相吻合这类推论所必需依凭的超基因遗传关系组;以及给定一种亲代的基因组排列,便可以据之预测出(经配子和授精过程而产生的)子代基因组排列的(孟德尔和摩尔根提出的)那些遗传规则。基因组和表型都是状态变量;因此,群体遗传理论把一系列基因组映射于一系列的表型,再把这些表型改变成其他的表型,然后再映射这种结果到基因组,最后将其转变为下一代的基因组排列。

Elliott Sober 在这一结构的基础上把进化理论构建成了一种作用力理论,亦即对哈迪-温伯格等式(Hardy-Weinberg equation)所规定的那种零外力均衡状态上的作用力理论。(哈迪-温伯格等式表明,在种群的第一代之后,除非受到外力作用,则每个基因座(locus)** 上等位基因(alleles)*** 的比率仍将维持不变,并且该等式还用一个公式表示出了这一比率。进化理论确定了,在各种外力(选择、突变、迁徙、基因漂变[genetic drift]等)的单独或联合作用下,这种均衡是如何改变的。(参见 Elliott Sober, *The Nature of Selection* [Cambridge, Mass.: M. I. T. Press, 1984], ch. 1。)然而,正如 John Beatty 指出的,哈迪-温伯格法则是孟德尔式遗传的一个结果,而且这种机制——有性生殖,雌性-雄性或雄性-雌性交叉生殖具有同等结果,满足分离规则**** 与独立配对规则的机制——本身就是进化过程的一个产物。如果进化理论也要解释孟德尔式遗传是如何发生的,那么 Sober 对此的论说就不可能是完整的了。(参见 John Beatty, "What's Wrong with the Received View of Evolutionalry Theory?" 载于 *Proceedings of the P. S. A., 1980*, ed. Peter Asquith and Ronald Giere [East Lansing, Mich.: Philosophy of Science Assciation, 1980], vol. 2。)

如果我们把进化理论看作是一种描述了不同零外力均衡状态的一系列历史性理论,那么,我们就可以避免这些困难来概括 Sober 的论说了。每种均衡状态都与一种遗传机制相匹配(在第一代中?),每种均衡状态的理论都确定了那种能够扰乱这种均衡的外力以及这种扰乱本身所具有的规则。有些扰乱将导致出现一种新的遗传机制,并且,一旦存在,这些新机制将生成自己的新的零外力状态,生成对这种新状态的扰乱规则等。由此,我们便得到了对一系列零外力状态及其相关机制的一种历史性叙说,亦即每一种状态与机制都依据与它们相关联的转换规则而产生出下一种状态与机制。对于每一种新的均衡状态而言,都会有一组新的偏离力(转下页)

不在于实际的繁殖成功——突发事故就能影响到这点——因此米尔斯(Susan Mills)和比提(John Beatty)一直主张，适合度乃在于生物生存和繁殖的(概率)倾向(propensity)。①

我建议把更大的适合度特质看作是一种存在量化(existentially quantified)表述。说生物A比生物B更适合于环境E，这就是说，存在着这样一种可遗传的(诸)表型(phenotypic)特征F,且F(作为原因)解释了在E中为什么A比B获得了更大的繁殖成功。因此说A比B在繁殖上更成功乃是因为A比B有更大的适合性，这话就不是同义反复，即使它说在环境E中A比B获得了更大的繁殖成功是因为存在着某种可遗传的表型特征F，它还是可以解释这种更大的繁殖成功。也许还有其他的方式能够解释这种更大的繁殖成功，比如说，偶然性。

存在量化(existential quantification)着重关注的是居间层次的表型特征，以及这些特征所完成的对于有生命力的后代之

(接上页)量的和它们是如何起作用的新规则。因此，在 Lewontin 的图式(schema)中，一种转换的输出结果就可以是一种新的自然(即零外力的)状态，具有不同的偏离力量及运作规则。然而，即便均衡状态、特定的遗传机制，以及偏离力量都是可以变化的——在这个意义上看，这种理论是极端历史主义的(radically historical)——但是自始至终，使其成为一种进化论故事的，乃是在适应性上的遗传变化所发挥的作用。

* 超基因遗传是指 DNA 序列不发生变化，但基因表达却发生了可遗传的改变，即基因组(geno)未发生变化而表型却发生了改变。——译者

** 在一条染色体上某种给定基因所占的位置。——译者

*** 在一个特定染色体上占据特定位置的一对或一组基因。——译者

**** 分离指成对等位基因的分离，尤指在减数分裂过程中的分离，结果为每对等位基因的成员出现在不同的配子中。——译者

① Susan Mills and John Beatty, "the Propensity Interpretation of Fitness", 重刊于 *Conceptual Issues in Evolutionary Biology*, ed. Elliott Sober (Cambridge, Mass.: M. I. T. Press, 1984), pp. 36-57。

存活与繁衍不可或缺的那些活动与功能。表型特征 F 总是通过更好更有效率地完成某项居间的一般性功能 G(比如逃避捕食者、寻找食物、能量转化、吸引配偶、获得足够的水分、保持热量等等)而实现这一功能的。列举出所有这些居间功能本身就会给一种适合度理论填充进更为具体的内容;假如做不到这一点,那么我们能列举出的居间功能越多,我们赋予适合度理论的确定性内容也就越多。适合度定义也许能参考这个居间(亦即居于表型特征 F 与繁殖成功之间的)层面的活动与功能 G,在这个层面上每项活动和功能 Gi 都是这样的:其他情况相同,该生物对 Gi 完成得越好,则它实现繁殖功能的概率也就越大。因此,说生物 A 比生物 B 更适合于环境 E,也就是说存在着某种(或某些)可遗传的表型特征 F 和某种居间层面上的功能 Gi,而 F 提高了 Gi 的表现,使得 A 在环境 E 中比 B(或许是概率性地)获得了更大的繁殖成功。 115

这里还有我们必须面对的另外一种复杂情形。(毫无疑问,为了对付更复杂的情形,我们需要添补更多的细节。)既然由于偶发的死亡、突变或者其他的因素,更适合的生物不需要有更大的实际生殖成功,因此,或许我们应该反过来这样说:如果出现了更大的实际繁殖成功,那么那种具有更大适合度的生物,其可遗传的表型特征将能够解释这种成功。在此情况下,A 比 B 更适合于环境 E 可以做出如下定义。存在某种可遗传的表型特征 F 和某种居间功能 Gi,以致:(a) 在 E 中 A 比 B 获得了更大的繁殖成功,而且是 F 提高了 Gi 的表现并因而(或许概率性)解释了这一点;或者(b) 在 E 中 B 相对于 A 有着更大的或者同等的繁殖成功,且不存在某种可遗传的表型特征 F'能够解释这一点,而且如果 A 在 E 中有比 B 更大的繁殖成功,那么这种成功便可以用 A 具有的这种可遗传特征 F 来加

以解释。

然而,因为这些条件会受制于复杂的反例,所以也许更好的是避免这些反事实的复杂性与定义性条款。我们可以反过来认为,只有实际上存在着更大的繁殖成功时,才存在着更大的适合度,尽管这也可能是其他原因所引起的。这可以被(粗略地)说成,更大的适合度是(概率性地)由可遗传的表型特征(通过某些居间功能的作用)所引起的更大的繁殖成功。没有实际的更大的繁殖成功,便没有所谓更大的适合度。然而,那些**确实**展示出了更大繁殖成功的其他生物,并不能被自动地算作更适合的生物,因为尽管它们有更大的繁殖成功,但这并不一定是由某种可遗传的表型特征(概率性地)引致的。(尽管这种建构将使最适者生存成为同义反复的,但是生存者的适合度却并不是。因此,我们可以加上的一个经验性观点是:就绝大部分情况而言,幸存者**就是**更适合的。)

我们所描述的适合度概念是一种比较性概念("在环境 E 中比_更适合")。出于多种目的,人们想要一种比作比较更强的适合度测度。这个测度不一定要是单个实数,而可以是那种更复杂的数学实体——一个有序的 n 维、一个向量、一个矩阵、一个树型的矩阵结构或任何单位。比提(John Beatty)和芬森(Susan Finsen)指出,生物有何种概率在其子代中确切地留下 n 个后代,这本身就是该生物的一个繁殖策略问题,而且在这类策略中可以存在着选择。① 我认为我们可以反过来考虑这样一个向量[$p0$、$p1$、$p2$、……、pn],这里 pi 表示生物在其子代中确切

116

① 参见 John Beatty and Susan Finsen, "Rethinking the Propensity Interpretation", 载于 *What the Philosophy of Biology Is: Essays for David Hull*, ed. Michael Ruse (Dordrecht, Kluwer, 1989), pp. 17-30。

留下 n 个后代的概率(并且所有 i 都大于 n 的概率为 0)。为什么要抛弃这个信息的任何东西呢?[①] 除了第 0 代的向量以外,我们想要考虑的是一个生物的更长时期的适合度,即它在第二代中确切留下 i 个后代的各种概率,这种适合度将是与其(潜在的)每个后代的一代向量(one-generation vectors)跟随其亲代的那代向量的一种函数;以此类推也是后继各代向量的一种函数。但是这里有一个难题。对这些数学结构的具体规定将是各种生活历史的(life-history)特征,而其间也是会存在选择的。那么,我们能够依据**哪种**适合度概念来解释 A 在环境 E 中比 B 有更大的适合度呢?如果一列向量占优于另一列向量,那么问题就简单了;但在不存在占优关系的情况下,复杂性就随处可见了。实际的历史过程最终有可能是**不大可能**的那条路径,且甚至连最适者**很可能**(*probable*)生存(如我们最初的叙述所解释的那样)这一概念本身也没有清楚的意义了。因为所使用的特定适合观要取决于有待解释的特定现象,这个一般性难题由此可以得以避免吗?

我们所知道的进化过程不只包括适合度上的可遗传变异,而且也包括遗传物质在后代中的不完善复制。绝大多数突变都是有害的,而且(比较复杂的)生物拥有一些修改及更正错误复制的机制;但这些修正机制是不完善的。即使这种完善是可能

---

① Robert Brandon 发现有影响的不只是对于下一代之数量的预期值——越来越大的方差(variance)也对选择不利——因此,他建议从对后代的预期数量中减去方差的某种作用后再测量适合度。参见 Robert Brandon, *Adaptation and Environment* (Princeton: Princeton Univ. Press, 1990), pp. 39 – 77。但要减掉**哪**一个作用呢? Beatty 与 Finsen 认为这并不是一个简单的均值与方差的问题;分布的扭曲也很重要。因为特定的统计数据只是该生物在一个环境中所采取的策略的一个成分,因此,一般意义上的适应性就不能等同于任何一个统计。参见 Beatty and Finsen, "Rethinking the Propensity Interpretation", pp. 17 – 30。

达到的，那么它也很有可能得不到选择。我们都是这种突变所衍生的后代，它们与那些能更精确地复制的同类的竞争中干得很不错。然而，若一个太挑剔的纠错机制来掌控复制过程的话，它将无法保持它所允许的那些突变，也无法在代代沿袭的过程中保留**它自身**——它是一种继承的装置。因此，基因纠错机制的准确度问题本身就是受制于选择的，我们实际的普通机制所容许的那种松动余地不应被视为一种缺陷——那些具有更完善纠错机制的生物还是原生动物。[①]

117 除了比较不同的生物繁殖成功的适合度概念之外，我也相信还有更一般的适合度概念的存在空间。我们来考虑这个问题，究竟为什么会存在生殖实体（或生物）呢？一旦生殖生物存在，它们就会繁衍扩散。是否应当有一种更为一般的适合观，以囊括那些生殖性生命体相对于无生命体所具有的优势呢？或许这种衡量也应当谈及对原子或分子的竞争：活的生物是否成功地将它们整合进了生物之中呢？因此，在局域的（封闭的）物质环境中，我们就能得到生物量与非生物量的比例。［列旺廷（Richard Lewontin）告诉我，在植物学中，生物的**数量**并不总是清楚或者重要的，因此，在植物学中，一种聚焦于其他材料（material）测量的适合观可能也是有用的。］有多少宇宙物质能够被整合到生物量当中去，这存在理论上的极限吗？在地球范围内，我们是否仍然处于生物量的上升曲线上呢？这条曲线在

---

① 也许把基因纠错机制与一台图灵机做比较是有用的。后者有扫描器会逐格扫描，一有必要就做出改变，即从周围材料中给磁带加上新格子。或许计算机理论的有些形式结构与结论能阐明这一生物学部分吗？或许我们需要一种稍有缺陷的图灵机理论，它有时候会算错。请注意一个纠错机制修补现象中的自我指涉性模拟（self-reference analog），除了其他的材料以外，这个修补因果地负责**该**纠错机制自身的（下一次）复制。

不同时期呈现出的又是何种形状呢？当我们构建一种更为一般的概念来阐明繁殖生命体为什么能存活和扩散时——通常的那种生物适合观作为一种具体的规定而被包括进来——又会引出许多新奇且有趣的问题。

我们一直把合理性视为一种带有某一功能的生物性适应。什么是一种功能呢？功能观是如何契合于一种生物学与进化论框架的呢？内格尔(Ernest Nagel)为我们提供了一种富有启发性的自我平衡系统(homeostatic system)的分析，比如我们身体内的温度调节系统。在某种环境中，这种系统会把某一个状态变量的值 V 维持在某一范围之内，当 V 被引致与此范围偏离开一定距离(不是任意大的一个距离)的时候，其他变量的值就会来补偿 V，通过修改这些变量值以使 V 能够回到原来规定的范围之内。[①] 内格尔将此作为对一种目的论或目标导向(goal-directed)系统的分析，该系统的目标或功能就是把变量 V 维持在该范围之内。按照这一论说，任何一个与 V 具有普遍联系的其他变量 V' 也将构成这种自我平衡系统的"目标状态"(goal-state)，查看为什么该系统要把 V 维持在那一范围内的解释，也许可以避免其反直觉的结果。然而，并非带有一种功能的每件东西都是一个自我平衡系统。餐厅椅的功能或许是支撑一个坐在桌边的人，但假如这把椅子没能很好地完成这一功能，它并不会去修改自己的一些状态变量以便更好地支撑人的身体。 118

为了应对这样或那样的情形，赖特(Larry Wright)提出，某物的功能有助于解释它为什么存在：当 Z 是 X 存在的一种后

① 参见 Ernest Nagel, *The Structure of Science* (New York: Harcourt, Brace and World, 1961), pp. 401 - 428; Nagel 遵循了生物学家 G. Sommerhoff, *Analytical Biology* (London, 1950)。

果或结果，并且 X 的存在乃是因为它能完成 Z 时，那么 X 的功能是 Z。[①] 波亚斯(Christopher Boorse)反驳道：如果一位科学家制作了一条带有裂口的管子，并且在试图将裂口修补好之前，他便被从裂口处泄露出来的煤气给毒死了，那么，在这种情况下，泄露煤气并不能算是裂口的一种功能，即使裂口导致了漏气并且是因为漏气而继续存在。[②]

让我们重新审视。若 Z 是 X 的一种后果，并且 X 就是为了获得这个效果 Z 而被设计或塑造(或维持)成这样的，那么 Z 便是 X 的功能。而这种设计或塑造，要么是出于某个人类设计者之手，要么便是进化过程的产物。无论是哪种情况，这个设计本身似乎便是一种以 X 能产生 Z 为目标的自我平衡过程。一位人类设计者构建了一把椅子以便其能支撑一个人，并同时改变椅子的某些特征——在设计或实际制作过程中——以便可以有效地实现这种支撑功能。一代代以来，进化塑造了生物体和身体器官，以能更有效地达到某些效果；并且，进化选择了那些能达到这一点的生物。[当然，除了适应性选择以外，其他过程——比如基因漂变(genetic drift)——也在进化中起着作用。并且，把进化在很大程度上看作一种自我平衡机制也没有蕴含它是最优的机制。]我认为，我们可以组合内格尔和赖特的观点，从而提出一幅更为完整的功能画面：若 Z 是 X 的一种后果(效果、结果、特性)，且 X 产生 Z 本身乃是某种满足内格尔分析的自我平衡机制 M 的目标状态，并且 X 是由这一自我平衡机制 M(通过对 X 产生 Z 这一目标的追求)所产生出来或维持的，那

---

① Larry Wright, "Functions", *Philosophical Review* 82 (1973), 重刊于 *Conceptual Issues in Evolutionary Biology*, ed. Sober, pp. 347 - 368。

② Christopher Boorse, "Wright on Functions", 重刊于 *Conceptual Issues in Evolutionary Biology*, ed. Sober, pp. 369 - 385。

么 Z 便是 X 的功能。(人们也可以去掉要求 X 实际上产生了效果 Z 的第一条款,因为并非所有的功能都得到了实现。)

这种论说解释了为什么生物学家不会说垃圾 DNA 的功能就是啥事不做或者就是要成为不值一除的东西;也不会说偏分离(Segregation Distorter)的功能就是破坏减数分裂(meiosis)。[①] 尽管这些都是效果,但垃圾 DNA 和偏分离不是由一个具有这些效果的自我平衡过程所塑造的。没有任何自我平衡过程旨在产生这些效果或是因为这些效果而得到选择的。要注意这里提出的观点并不自动地把一种自我平衡系统的目标状态作为其功能。一些部件的偶然组合创造的恒温器并不具有调节温度的这种功能,即便它能产生这种效果。在我的论说中,自我平衡系统本身就是设计者,而不是被设计的对象,而且其目标就是某物 X 产生效果 Z,这个 X 是系统所创造或维持的,并且这一功能正是要归因于 X 的。那么 X 的何种效果是其功能呢? 答案是,自我平衡机制创造(或者维持)X 而想达到的那些效果就是 X 的功能。[由于某些副效果与那些得到选择的效果是共外延的,因此要么功能的归因必须附系于为什么该效果得到选择的一种解释。(这里的"解释"不是一个外延[extensional]概念),要么这里提出的论说只是一种效果成为一种功能的必要条件而非充分条件。]

119

进化过程产生了带有各种功能的实体,但是只有当存在着其他某种(产生或维持进化过程以保有该效果的)自我平衡系统时,进化过程自身才会具有一种功能——比如说,创造出这类带有诸多功能的实体。要注意这并不是这样一种观点的产物,即

① 这些例子来自 Peter Godfrey-Smith,但对于为什么这些不是功能,他给出了另一种说法。

事物(在道德意义上)*应当*完成它们的功能;如果某种存在物是为一种自我平衡系统所创造、塑造而成奴隶的,且要拼命劳作,那么这种存在物(基于我的论说)就具有这一功能,然而尽管如此,他们仍然应当反抗。还要注意到,不是由一种自我平衡系统所塑造成的某种存在物,也有可能会在某个时间点上开始被用来为一种目标服务——比如说,一块平坦的大石头被用作野餐桌。对此,我们也许会说它已被*赋予了*一种功能;我们可以用它和那些通过一种自我平衡机制而保有这种功能的事物做同样的事情。它有一张餐桌的*功能*,尽管这并不是它的功能。但如果我们现在要塑造或维持这个对象,以使它保持在达到该效果所需的那些特性的范围之内——比如说,清理掉那块大石头上面的苔藓及落枝——那么,保有那种效果的持存状况——作为一张野餐桌的可用性——便成了一种自我平衡机制的结果,而且在这种情况下也就*变成*该机制的功能了。

合理性的功能

对于理由所支持的东西而言,这些理由本身就是证据。[①]但是我们为什么要出于理由相信和做事呢?*理由的*功能是什么?这个问题的答案看来是显而易见的。理由是与它们所支持之物的真理——亦即事实性关联——相联系的,因此,出于理由而相信是通往相信真理之道。尽管如此,这一问题仍然值得做更深入的探究。根据我们对功能所作的分析,"我们出于理由相信或行动"的功能就是它具有的某种特性或效果,而这是背后的

---

① 我曾表示,这存在于一种双重关联中:既是一种事实关联,也是一种结构性关系,并且看上去是自明的。

某种自我平衡机制“旨在”让它具有的。既然合理性是考虑理由(并据之行事)的，那么合理性是什么(即合理性有什么功能)就将取决于这个世界的一个事实，亦即在塑造我们基于理由而相信和行动中，(诸)自我平衡机制实际上是如何运作的，它要达到什么样的目标。 120

要考虑的第一种自我平衡机制就是通过自然选择来运作的进化过程。出于理由而相信和行动是一种被选择的特质吗？如果是的话，为什么呢？请注意，那种出于理由而相信和行动的特质，可能不是作为直接目标而只是作为一个**副产品**被选择的。在此种情况下，出于理由相信和行动就不是一种进化的功能；不存在任何这样的属性 P，以致一种进化的自我平衡机制会把出于理由而相信和行动具有属性 P 这一事实作为其目标状态。我们已说过，正因为世界以不规则的方式变化，所以生物才需要那些适应机制以应对当地环境；整个工作无法在操作性条件反射下，由一劳永逸的结构和预配反应而有效地完成。理由及推理都有助于生物面对新的环境和避免未来的麻烦。无论合理性的这样一种能力是被直接选择的，还是其他获选能力的附带产物，它都将很好地服务于生物的生物使命，并提高它的全面适合度。这将赋予外显的合理性去应对变动的事实和变化的需要这一任务，而且当我们发现当下的变故使得我们的种系发展(phylogenetic)行为模式不再适应时，也许就要修正它。

进化可能给我们灌输了有关我们进化史的稳定事实的种系发展信息，还灌输了如何适应它们的行为模式。我们没有必要明确地思考(或甚至知道)日夜的规律交替；我们的生理节奏一直在替我们做这个工作，正如感受到了时差的旅行者们会发觉的那样。有些事实是如此稳定，以至于任何生物自身都没有必要意识到它们的存在；进化将它们植入进来了。合理性可能拥

有这种进化功能,使得生物能更好地应对新的变动不居的当下情况,更好地应对在一些(或许是复杂的)当下迹象中所预示出的未来情况。合理性能这样做本身就是那些稳定的事实之一,而不是需要我们去明确地知道的一个事实。正如重力的出现一样,我们要做的一切就是让它嵌入我们。进化是围绕稳定不变的环境来建立且运用各种机制的。重力不仅是对一些特性的限制——比如说,大小——重力本身也在一些过程中得到了利用。过久的失重状态会使宇航员的生理状况受到严重损坏,这是因为他们的生理机能是在一个稳定的重力环境下进化而来的,并且相配而得到利用和起作用的。这些过程不会再重复重力已经做过的事情。(在有重力的情况下,重复它的企图只会产生过度的作用力,导致有害的结果。)

121

让我们来检验下面的假说。长期困扰着思想家们且未得明显成功解决的哲学难题之清单——如归纳问题、他心问题、外部世界之存在的问题、理性的证成问题——全都标示着进化植进我们的那些假设、还有其他的但不那么常见的问题。

但我们为什么认为有必要去解决这些问题呢?如果迄今为止的所有人都降生于由其他人所组成的环境中,那么,他们用不着亲自去学习或要推理出:其他人与他有着类似的心灵。每个人要有效地发挥作用就必须知道这一点;我们没有理解这一点的那些远亲就不会留下类似的(同样不能知悉这一点的)后代。那些最快掌握了这一点的人们就拥有某种优势,进化过程因而(通过鲍德文效应)给我们植入越来越容易的习知能力,直到最终这一知识嵌入我们之中。类似地,那些无法获知存在着一个独立的"外在世界"(它们即使没有被观察到也会继续处于其轨道或位置上)的远古亲戚,也不会如那些能快速认识到冷酷实在的人们过得那么好。那些不能学会从既往经验中(以对当时而

言适当的方式)做总结的人将遇到各种危险,由此留下的后代也会屈指可数。合理性的功能从来就不是**证成**这些体现了我们进化史之稳定性的各种假设,而是在这些假设确立的稳定框架**之内**,利用这些假设应对变动不居的环境以及各种问题。因此,我们的合理性工具无法为这些假设提供结论性的理由或“证成”,这便是不足为奇的事情了。它们并非是为了那一目的或为了给**自己的**运用提供结论性理由的目的而被设计出来的。[①]

巴特勒主教(Bishop Butler)说,“概率是生活的向导。”但在具体的场合下,我们并不能基于什么是最有可能的而确立行动的合理性。实际上,直到最近,依据绝大部分现存概率理论而言,将一项概率归属于某一个特殊事件都是没有意义的。我们需要给出理由来表明,例如,为什么在一种假设的无限事件序列中会发生的事,应该用来在某一特定的非常有限的情形之下指导我们的行为。最近,有人构建了概率的各种倾向诠释(propensity interpretations of probability)理论来把各种概率归于特定事件。但尽管如此,依据这些理论且依据把概率视为在科学内用来解说特定现象的一种**理论**术语的这一观点,我们仍然没有解决这个问题:为什么我们下一次还应当依据最大概率来行事呢?在效用理论中,依据概率而行事的合理性由冯·诺伊曼-摩根斯坦条件所表达,即如果个人偏好 x 于 y,那么在只产生有不同概率的 x 与 y 的两种概率混合物中,他会偏好 x 有着更高概率的那个混合物。如果个人偏好 x 于 y,则[当且仅当 p 大于 q 时,有 $px, (1-p)y$ 优于 $qx, (1-q)y$]。这一表述 122

① 这未必意味着:我们曾经更善于发现及证明(例如)那些归纳的原则、他人的思维及外部世界的存在,且我们自那时便特意地配合着这些事实而行事,而这是我们现在再也无法证成或证明的了。

与卡尔纳普还原语句具有的形式完全一样，[1]由此暗暗地表示了以这种术语来定义概率的一种方案。这里不是去操心如何证成我们为什么应当依据概率行事，相反是要考虑根据我们应该如何行动来定义概率。萨维奇(L. J. Savage)实行了这一方案，他对行为，即对行动的偏好设定了一系列的规范(及结构性)条件，从而足够来定义一种个人概率的概念。[2] 例如，如果一个偏好 x 于 y 的人选择了 A 行为而不是 B 行为，那里当处于状态 S 时，A 产生 x，当处于状态 T 时，A 产生 y，而 B 与此相反，当处于状态 S 时 B 产生 y，处于状态 T 时产生 x，那么，我们便可以把他选择 A 行为看作表明他的看法是，正如对他的看法的定义，S 比 T 的概率更高。但是，依据在行为中的这样一种偏好来定义概率也会遇到麻烦。假设这同一个人同时还偏好 z 于 w，并且选择了 C 行为而非 D 行为，那里 C 在状态 T 时，产生 z，在状态 S 时产生 w，D 与此相反，在状态 S 时产生 w，在状态 T 时产生 z。这将表明此人认为 T 比 S 的概率更大。但是，他早前偏好行为 A 而不是 B，这表明他认为 S 比 T 的概率更大。因此，如果谁想要依据个人在行为上的偏好来界定个人概率的话，那么他就必须避免这种冲突。萨维奇因此加了一个条件：如果这个人如上面那样对于具体的结果 x 和 y 而言偏好 A 于 B 的话，那么，对于每一个 z 和每一个 w 而言，只要该人偏好 z 于 w，那么她在上面都将偏好 D 于 C。然而，并不存在任何独立的理据将此作为一项规范要求，除非该理据承认某种独立的而不是个人性的概率概念，还要把个人的选择看作是标明她是按照

---

① Rudolf Carnap, "Testability and Meaing", *Phiolosphy of Science* 3 (1936): 419 - 471 and 4 (1937): 1 - 45.

② L. J Savage, *The Foundations of Statistics* (New York: John Wiley, 1954).

她相信是更大的概率而行动。[①]（她也许反而是按照根本就不涉及概率的原则行动的，或者是以不同的方式涉及概率而行动的。）因此，这个条件假定了个人始终应该具有某些明确地规定的概率信念，并且总是会（以冯·诺伊曼-摩根斯坦公理所规定的方式）依据它们而行事。但有争议的正是这一点：为什么依据更大概率来行事就是合理的呢？萨维奇试图根据行为来定义概率的尝试也无法避免这个问题。[②] 然而，工具合理性观念与信念的合理性之中的核心问题是：为什么在具体的情况下，我们应当按照最大的概率来行事，或相信最可能的事情将会实际上发生呢？迄今为止所具有的合理性资源对此都无法提供一个令人满意的答案。[③]

123

---

① 我接下去会讨论荷兰赌论证，它并没有提供出这样的一种独立的理据。它最多表明了：如果存在着一个人们总会遵循的个人概率，那么它们为什么应该满足于那些通常的概率公理；它并不是表明，为什么必须或者应当存在这种总是指导人们（下赌）抉择的个人概率。

② 尽管我已经提出了这段中的具体观点，参见拙著 *The Normative Theory of Individual Choice* (1963, rpt. New York: Garland Press, 1990)，然而，只是在我与 Hilary Putnam 交谈之后，并且在我读了他的一篇最近的未刊文章"Pragmatism and Moral Objectivity"（即出）之后（他在文章里独立地提出了个人为什么应该总是依据更大可能性来行事），我才开始把这个问题当作一个严肃的问题，而不仅仅是当作一种刁难。（另一个相关的问题，为什么个人应当按照确定性而不是按可能性方式行事，**也许**可以为占优考虑所解决，**如果**这样的考虑没有回避问题实质的话。）

③ 最后，我们考虑一个合理性自身的问题。设陈述 R 为：当且仅当陈述 p（或行为 A）能被表明为合理的时候，我们才相信 p（或做 A）。（或者当你最初获得这种信念的时候，它本来能够被表明是合理的。至于维持这种信念，只要它不被表明为不合理的即可。）我们拥有强大的归纳根据——亦即，好的理由——来假定 R 本身不可能被表明为合理的——尽管做出了各种各样的严肃性努力，但尚无人做到这一点。假定情形就是如此。那么如果 R 是真的，既然它不能被表明为合理的，你也不应当相信它。那么，至少有一种真理是合理性所无法带给你的。（若那个真理是这样的，为什么其他的真理不是这样的呢？）另外一方面，如果 R 是假的，那么就存在着有些你相信或做的事情，尽管不能表明它为合理的，**或者**，存在着某些你不会相信或做的事情，尽管能够表明它为合理的。在上述任何一种情况之下，合理性（转下页）

我们来考虑康德的这一尝试,即把有原则的行为作为行为的唯一终极标准,而不考虑个人可能具有的任何特殊欲望。但是,原则乃是为了与人们的既存愿望配套而形成的装置——我不是说它们是一种进化适应物——其中有些原则是生物性地植入的(biologically instilled)。因此,原则是一些局部装置。然而,康德通过剔除与原则一起起作用的共同因素而裁掉了原则能够在其中起作用的环境;只让原则(也就是原则这一概念本身)独自来完成所有的工作。我们前面在讨论原则的章节中,可没有赋予原则如此空洞的范围或任务。我们的合理性及我们的原则不仅是局部的,而且是设计来与外部事务相配套起作用的。而且我们人类本身就是局部性生物,并不是完全自主的。我们是自然界的组成部分,设计成与自然界的其他部分、事实相配套而起作用,且依赖于它们。人类的记忆利用储存在(有序)外界事物中的信息;[①]我们同样还是占据着某些生态位(ecological niches)的物理造物。对于"合理性"及其局限的这种进化论说契合于维特根斯坦、约翰·杜威、马丁·海德格尔以及迈克·博兰尼(Michael Polanyi)等人著作中的一种论调:他们同样(出于不同的理由)也将合理性视作嵌入在某一语境之中并且作为众多成分之一来一起起作用的,而绝不是用以评判所有事物的一个外

(接上页)看来都是有局限的。有很强的归纳根据让我们相信:一旦规定任何一种特定的合理性观,R 便不可能被表明为合理的。根据合理性自身的标准,它(迄今)都无法证成它自身。那么就我们相信 R 不能被表明为合理的而言,这看来是合理的,因此如果 R 是真的,不要相信它。(当规则 1 和 2′代替 R 时,情境是如何变化呢?)

① 参见 Donald Norman, *The Psychology of Everyday Things* (New York: Basic Books, 1988), ch. 3。

在的自足观点。①

为什么我们不能用合理性证成某种假设，对此的进化解释——亦即我们的合理性是被设计成与这些事实一起来发挥其他功能的——本身并不是对于这些假设的一种证成。正如欧几里得几何学只需"足够真"那样，对于他心、独立存在的外部世界的信念也都可以(通过鲍德文效应)固定下来，不必要求其在严格意义上谈及真——这些信念所要的是"足够真"。这样一来，或许我们无法确信我们与之配套起作用的那些"真理"究竟是什么。

休谟进一步告诉我们，一项在过去成立的规则性并不能确保其在将来(很可能)成立。过去的事实导致一些与这些事实相符的假设被植入我们之中，这一点并不意味着过去那些事实将继续成立，也不意味着那些假设将继续为我们服务。纳尔逊·古德曼(Nelson Goodman)向我们指出，即便过去成立的某项规则将继续成立，但是还有许多与过去相符合的规则性在将来是会出现偏差的。何种规则性会继续成立呢？② 进化只是选择了那些迄今为止一直有用的特质。迄今为止同等有用的特征便会受到进化的同等喜爱，无论这些特征之有用性在将来会发生多大的分歧。进化并不特别偏爱未来的有用性，尽管在现在变成

124

① Hubert Dreyfus 论证，人工智能计划遭遇到的困难归因于人类合理性的嵌入性(embedded)和体现性(embodied)本质。参见他的 *What Computers Can't Do: A Critique of Artificial Reason* (New York: Haper and Row, 1972)。

② 这个问题是由古德曼讨论"绿蓝"(grue)与"蓝绿"(bleen)的作品迫使我们面对的，这两个不同寻常的谓说(predicate)在既往的时间内都符合绿色事物与蓝色事物之正常表现，但在未来却偏离了正常的谓说。参见 Goodman, "A Query on Confirmation", *Journal of Philosophy* 43 (1946): 383 - 385, and *Fact, Fiction and Forecast*, pp. 73 - 83。通过画出经过相同的过去数据点的不同曲线，也可以提出相同的观点。那么何种规则性会继续成立呢？

过去(先前它是未来)后,它会偏爱于那些表明为有用的特征。从这一视角来看,当人们声称是进化把过去的稳定规律灌注入了我们的种系发展遗传中时,问题就不仅是在于过去的规则是否会继续成立,我们的遗传特征能否继续为我们提供服务,而且是在于进化是挑选出了"正确的"规律,还是在一个"绿蓝"(grue)世界里却给了我们"绿色"世界。[古德曼的投射规则(rules of projection)能被解释成比较性适合度(comparative fitness)的标准吗?]

尽管如此,合理性不是设计出来证成自身或其框架性假设的,这一事实并没有蕴含合理性**不能**这样做。[然而,哲学家们迄今为止的败绩支撑了这一观点:把这种证成(甚或是这样一种推理)留给单个人个别地去完成是很没效率的。]这一事实也并不意味着,我们无法找到理由来**反对**那些框架假设具有不受限制的真理性,即便它们是以进化的方式植入我们的。回想一下欧氏几何学的例子,即便它是作为"自明"而被选择的,人们亦能发现它不是严格准确的。的确,我们可能是借助于进化植入进我们的其他各种理性关系(reason relations)——或者更切题地说,是借助于对那些理性关系所作的修正(这里引导我们做出修正的还有其他的理性关系)——才得以发现这一点的。我们没有必要从这种完全准确地反映了事实成立的那种理由来开始。一组**大致**(roughly)准确的理性关系能把自身塑造成一个更准确的关系组。一个大致准确的工具能够发现并修正另一个工具的缺陷;这第二个得到改进的工具又能够找出并修正下一个工具的缺陷;而第三个工具此时又可以对第一个工具做同样的工作。由此,所有的工具都能改进到一个新位置,这比它们开始时都更准确。长此以往,每个初始工具的性质都会发生重大的改变,并且这三个工具一起兴许还能设计出一个新的第四个

工具。①

我们还需要考虑另外一种自我平衡的塑造机制,即各个社会用来塑造其社会成员的那些过程。支持出于理由而相信和行动的那些能力或许一直就是自然选择的对象,无论理由是什么;而一俟这些能力存在,社会便抓住这一良机产生出(多少有点)理性的成员。当社会科学家们谈及合理选择时,他们的目的常常是为了解释社会制度的特征,并试图展示理性个体是如何组成和维持一个社会的。② 这

---

① 比照 Wittgenstein 在 *On Certainty* 中的如下话语:在第 83、88、94、103、105、152 段落中讨论"参照框架"(frame of reference)、"我作为立足点的命题"(Propositions that stand fast for me)以及那种"固着于我全部的问题与回答中,是如此之牢固以至于我无法触及"的东西,对比于我们的这一假说是有益的:即把历史环境中的稳定事实作为种群发展遗传而植入我们和进化假说。然而,Wittgenstein 没有考虑这样的方式,这些植入的框架成分以之操作可以引致彼此改变。

② 参见 Douglass North, *Institutions, Institutional Change and Economic Performance* (Cambridge: Cambridge Univ. Press, 1990); Andrew Schotter, *The Economic Theory of Social Institutions* (Cambridge: Cambridge Univ. Press, 1981); Oliver Williamson, *The Economic Institutions of Capitalism* (New York: Free Press, 1985); Margaret Levi, "A Logic of Institutional Change", in *The Limits of Rationality*, ed. Karen Schweers Cook and Margaret Levi (Chicago: Univ. of Chicago Press, 1990), pp. 383 - 401; Thrainn Eggertsson, *Economic Behavior and Institutions* (Cambridge: Cambridge Univ. Press, 1990); Harold Demsetz and Armen Alchian, "Production, Information Costs and Economic Organization", *American Economic Review* 62 (1972): 777 - 795; Michael Jensen and William Meckling, "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure", *Journal of Financial Economics* 3 (1976): 305 - 360 (rpt. in *Economic and Social Institutions*, ed. Karl Brunner [Boston: Martinus Nijhoff, 1979], pp. 163 - 231); E. Furbotn and S. Pejovich, "Property Rights and the Behavior of the Firm in a Socialist State", in *The Economics of Property Rights*, ed. Furbotn and Pejovich (Cambridge: Ballinger, 1974), pp. 227 - 251; Dennis Mueller, *Public Choice II* (Cambridge: Cambridge Univ. Press, 1989); Gary Becker, *The Economic Approach to Human Behavior* (Chicago: Univ. of Chicago Press, 1976); James Coleman, *Foundations of Social Theory* (Cambridge, Mass.: Harvard Univ. Press, 1990); Richard Swedberg, *Economics and Sociology* (Princeton: Princeton Univ. Press, 1990)。

一工作是富有启发性的，但我们还应当探究的是：社会如何以125 及为什么要产生并维护理性的成员。人并非天生就是理性的。不管有些合理过程在何种程度上是天生控制模式的产物，它们也还是要受到社会灌输的过程、规范和程序的塑造和影响的。那么是怎样的社会过程做了这种塑造，且为什么这些社会过程又存在呢？（仅仅是因为由社会灌入这些特征的基于理由而相信和行动所带来的因果效果吗？）

人们在其所面对的限制结构与激励结构中进行行动和选择，由此而维续社会制度和社会结构。这些限制结构和激励结构自身乃是由其他制度和其中的他人行为所确立起来的。这就像大型的拼图游戏那样，其中的每一片都只能填入由所有其他的图片所剩余下来的那块空间，每个人的行为也是在所有其他人的行为所留下的那些限制和激励空间内进行的。当制度能自我复制时——亦即当它在社会中与其他制度相结合时，它们能招募、训练新成员和公务人员，有能力为他们提供激励——那么这些制度就会具有连续性。部分训练可能涉及合理的规范与惯习、选择模式及信念的形成过程。① 制度为了在下一时期能够自我复制（或许会有点改变），就要创造出多少有点理性的个体，他们能够对特定类别的激励做出回应，学习并考虑特定类型的约束。我们不一定要把制度想像成是在试图自我复制的。这种过程是一个选择的结果：那些（即便与其他制度结合在一起）不传播、不再生产或不复制自身的制度将不复存在。这就是为什么我们能发现，几乎所有的现存制度都有手段来招募和训练新

---

① 此外，一种制度可以这样来发挥作用吗？亦即它好像是把制度内的个人行为联结起来以最大化某种客观功能那样，尽管这些个体根本没有打算最大化这种或那种功能。

成员，以便能接管维续这些制度所必要的那些功能。①

因此，合理性的一项显著功能或许是将制度传播到随后的制度阶段，而不是服务于那些得到训练而被塑造为理性者的利益。（制度所具有的确切性质将会影响到这种塑造过程是加强还是减弱理由与可靠地达致真理之间的关联。）在《自私的基因》（*The Selfish Gene*）一书中，理查德·道金斯（Richard Dawkins）把生物及其行为视作选来服务于基因繁殖的装置。②（“一只小鸡是一个鸡蛋用来制造另一个鸡蛋的方式。”）而我们这里的反思提出了另外一种可能性，亦即，合理性之所以被塑造、选择和维持，并不是服务于一个比生物更低的层次，而是更

126

① 除了合理性的社会功能外，还存在着合理性的社会特性（social character）主题，亦即在那种特征中合理性的方式是人际间的。Jurgen Habermas 认为：如果没有听取来自每一渠道之意见的开放意愿，如果这种潜在的意见来源没有必需的自由与资源去进行有理据的讨论的话，则合理性便是不可能的。但合理性并不要求进行最广泛的证据筛选和计算等。因为这些过程本身也会有成本，对某个具体的决策与信念形成要投入多少时间与精力，这也会有一个大概的决策。请注意，方才我提到的成本是属于个人性成本，而非特属于认识的成本。如果这些限制并不必然使任何特定个人的信念与行为变得不合理的话——实际上，对他们来说，不施加任何这类限制才是不合理的——那我们就不清楚：为什么说社会性的限制、在意见形成上欠缺完全的民主参与等必定会阻碍社会中的个人观点——或阻碍社会本身的意见，如果某人持有一种“社会合理性概念”（notion of social rationality）的话——成为合理的呢？可能有人主张，据之做出此种成本限制的那一过程的本质，才是问题的关键。然而，与个人自己决定准备承担多大的投入相提并论的是，一个非民主社会的统治者能够决定他们自己的信息来源和相反舆论的范围有多广；因此，这些人的信念与决策为什么就不可能是合理的，这是不清楚的。确实，他们也应考虑到，他们所使用的有限信息资源中可能带有偏见，但是，这种偏见能从更广泛意见的有限取样中得知。因此，贯穿整个社会的民主、公开辩论和意见形成过程，对于社会成员（或社会自身）形成合理决策与信念而言，似乎并非是必要的条件。古希腊并没有因为奴隶的存在而阻碍其社会成员形成合理的信念。民主权利具有不同的基础。

② 参见 Richard Dawkins, *The Selfish Gene* (Oxford: Oxford Univ. Press, 1976)。

高的层次，也就是制度层次。[1][2] 但这并不意味着我们可以忽略个人与制度之间的互动：每一方都以一种方式塑造着另一方，且其方式会影响到对自己未来的塑造。

选择究竟会在哪个层次运作？我们对此有时会提出相竞争的假说。我们考虑经济学家（标准地提出）的财富最大化假设。这一假设（比效用最大化更具体）为其理论赋予了更详尽的内容。“最大化”因为其在数学上的好处理和力度而得到援用，不过，我们打算考虑一种更弱但更可行的假设，即人们会很认真地对待财富（即便人们并不追求最大化且最大化不具有词典式第一的地位）。根据塑造人们的心理关切与动机的那种制度，根据特定的动机有助于这些制度的运作和传播的那种方式，我们或许能对此给出一种社会性解释。然而，还存在另一种可能性。据报道，各个社会中广泛存在一个现象，富人们倾向于拥有更多的孩子（尽管在近 150 年来的西方工业社会中并非如此）。在多偶制的社会中，富人们（通常是男性）倾向于拥有更多的配偶（妻子），而在所有的社会中，富有者都更有能力，能更好地为其后代抵御物质的匮乏。[3] 还有这样一项证据，即在 20 世纪的渔猎采集（hunter-gather）社会中（显然是符合大部分进化史的最近的

127

---

① 关于选择单位或选择在哪个层次发生的问题，生物学家与生物哲学家之间一直有着大量讨论。群体选择问题一直都激起这种讨论。一个经常被引证为支持群体或群间选择但看来不能通过个体选择得到解释的现象，那就是兔子中黏液瘤病毒（myxoma virus）毒性的抑制现象。但值得注意的是，如果存在弱化菌株毒性的修饰基因活动的话，也伴随有导致减少偏分离的表型效应，那么这种现象可以归入个体选择。

② 这指南美洲原产兔体内有一种黏液瘤病毒，这种病毒不大致病，但若感染给欧洲兔，死亡率几近百分之百。——译者

③ 这个主张是由 Gary Becker 提出的，见他的 *A Treatise on the Family* (Cambridge, Mass.: Harvard Univ. Press, 1981), p. 102，他引用了一些支持性文献。

当代人)，首领和富裕者往往后代更多。让我们假定，所有其他情况相同，致富愿望强的人往往会比致富愿望弱的人积聚更多的财富；也就是说，他们成功的可能性更大。如果重视财富的心理倾向(predisposition)是能基于基因遗传的——我仅仅是表述一种可能性——那么，有这种倾向的个体往往会产出更多的后代，他们能活到生育年龄之后。他们的后代也往往将具有这种倾向，由此积聚更多的财富，生育更多的孩子。由此，这种渴望财富且愿为之奋斗的遗传倾向就会得到选择。这种人的比例会一代代地不断增加。(不可能有超过50%的人比平均数更富裕，还有诸如社会流动性等各种因素，这些会如何影响财富最大化者在人口中所占的比重均衡呢？)由此，人们就可以对经济学家的假设(如果不是财富最大化，那么也是非常向往财富)给出一种进化解释。我们当中关注高雅事物的人是如此之少，这是因为关注高雅事物的祖先后代很少，而我们反过来则是那些关心物质财富者的后裔吗？当然，生物学的假设与制度性的假设并不是互斥的；这两类因素可以互动地产生一种现象。①

我在前面谈过，合理性是自觉的(self-conscious)，因为它旨

① 生物性因素也可以同其他社会性因素相配合，比如说父母对其孩子心理的有意塑造——这显然是一种普遍现象。父母这样做也许是为了子代的福利，或者是为了他们自己在互动上的便利，抑或是由于某些更大的制度把其塑造成这样的。但还有一种可能性值得一提：父母自身有这样一种得自遗传的倾向，使其想把子代塑造得与自己更为相像，亦即在孩子身上强化自己的那些心理特征。这种倾向会使子代得自遗传的(就父母亲与孩子共享的心理特征的)心理倾向得到增强，那么这种要强化与塑造的倾向便也有可能得到了选择，至少当它们与其他那些有助于提高全面适应度的遗传性特征相组合时会是这样。(如果这种使子代的心理匹配自己心理的遗传倾向果真存在过的话，那我们是否可以设想，其基因基础与其他某些重要的心理倾向之基础在染色体上有着紧密关联呢？)如果他们的本性引导着亲代这样来养育子代，以使与他们共享的(或有时是遗传自他们的)那些心理特征得到增强的话，那么，(先天)本性-(后天)培育的影响难以解开，这也就并不足为奇了。

128 在纠正它所获知的信息和推理程序之中的偏见。早前，我们也想知道合理性的功能是否只是在进化植入的假设框架内的一个有限功能，其中有些假设标出了我们所无法解决的哲学难题。我们是否应当把避免思想偏见也视为只具有一种有限的功能，只在社会基本偏见所组成的框架内部提出以及权衡各种理由与信息呢？我发现这个想法还真是令人困扰。我们纠正偏见的范围能有多大？我们是否应满足于社会所呈现给我们的关于信念与行动的选项范围，满足于社会所植入的(用来进一步裁剪这些选项的)价值范围，运用理由只从这一范围*之内*的那些选项中进行选择？或者我们还是应当从所有的正反理由开始，尽可能宽泛地思考这些理由，并且纠正我们所能察觉到的在信息的社会传递和社会对于理由的衡量与评价过程中所产生的任何偏见，然后，尽我们最大的能力在这一无偏见的基础上做出我们的决策？(如果各个选项在这一过程中打成了平手，那么，我们就可以让遗传所得的那些社会价值及社会假设来做决定。)

第一个方案即使不是独断的话，看来也过分乐观了；而第二个方案看来是合理的方案。我认同后一方案。但是考虑所有选项则效率又太低了。因此，如果社会呈现理由中的那种偏见代表着这些理由所应该具有的权重(亦即个人经过仔细且详尽的考量后会赞同的那种权重)，那么遵循社会的指导就是有效率的。社会偏见是否可堪比拟于贝叶斯的先验概率分布呢？

不同类型的适应契合于不同的变迁速率。如果每个人都要从零开始学做每一件事，并全凭自己之力来开掘每一点知识，那效率就太低了。我们(由进化植入)的基因遗产是在长时期内去匹配或回应这些时段内的恒常性(constancies)这一过程中塑造而成的。中等规模的物体长期存在着，持续不断地运动着，重力*则连续不断地发出一种源自地心的强大引力——这些恒常性对*

我们的进化史有着深远的影响。(另一种长期存在的恒常性是其他事物不断地变动着,我们由此也塑造得具有一种应对这种变化的机制,将其作为我们恒久的遗传天赋中的一部分。)我们的基因回应的是(迄今为止)历经千万年变化的东西。

有些人说,社会传统、制度和行为规则是适应那种更为迅速地变化(历经数代而不是亿万年)的要素。① 还有代代有变的要素——例如说,哪种职业技能是现代社会(非传统社会)最需要的技能。(这种速率可以更快,社会所需职业[意料之外地]在一生内就可能会起显著变化。)此外,还有一些适应是针对人的一生内那些缓慢变动的因素;这些包括行为模式及持久的个人纽带。然后,还有天天变或时时变的因素。我们的感觉器官就适宜发现在我们身上发生的变化,社会信息机制则适于带给我们远处的变化信息。"对不同周期的环境变化存在着不同的适应模式集。"②

129

---

① "[人的行为]对周遭一般环境的这种适应,产生于其对一些非经自己设计且经常不为自己明确所知的规则的遵守……我们的行为是由这样一些规则所指导的,它们适应于我们所生活的那种世界,亦即,与我们意识不到却又决定了我们行动之成功模式的那些环境。"这些规则"是经由一个选择过程在他所生活的那个社会中进化而来的,因此是数代人的经验产物",F. A. Hayek, *Law, Legislation and Liberty*, vol. 1: *Rules and Orders* (Chicago: Univ. of Chicago Press, 1973), pp. 11 - 12。哈耶克描述的这个过程是一种群体选择过程。"因此,这些行为规则并不是被当作为了达致一项已知目标的公认条件而获得发展的;它们是进化而来的,因为遵守了这些规则的群体获得了更大的成功并且取代了其他的群体。"(p. 18.)

② E. O. Wilson, *Sociobiology* (Cambridge, Mass.: Harvard Univ. Press, 1975), p. 145. Wilson 书中的第 7 章详尽地阐释了这一论点,他区分了对于不同变动周期的不同层次的回应:生物的、生态学的 (ecological)以及进化的。

针对每一层次的变动频率,都有一些适应性的机制来回应(近似于)该频率的变化,通过适当的反馈规则做出修正,它们创造了持续时间(大概)那么长的事物或实体。我们可以区分出如下情形:所造物与适应它要适应的那种恒常性(constancy);恒常性发生了改变,所造物也要做出改变来适应新的恒常性——正走向一轮新的均衡;如果恒常性以一种比回馈机制所能回应的速率更快地变动的话,那么新的所造物与当时的恒常性二者间不可能相适应。

有时候,因为社会中的其他人接受某事,你也接受它就是合理的。考虑这样一种信念机制,它使你接受(你能看到)大部分其他人也会接受的东西。我们都是可能犯错的,因此当我们与大家面对理解上差不多的一个问题时,许多也可能犯错的其他人所达成的共识便很可能比我自己的特定视角来得更为准确。就大量的情境而言,大量观察样本所得出的均值(mean)往往比一个随机选取的单独观测更为准确。(假定这些观测结果是围绕真值的正态分布,或者是由真值与随机误差因子所共同决定的。)那么,面对这类问题,你就应当纠正自己以便更接近于共识观,除非你有某种特殊的理由,认为其他人都受到了误导而你没有——例如,他们都以某种方式受到误导而你并没有。[1] 如果多数派观点是根据神话与迷信而形成的,而我的观点是以科学文献(或有关报告)为基础的(其中的观点是更仔细且更可靠地得出的),那么我便有理由认为少数派观点是正确的。马克思主义传统认为,其他人的观点是被意识形态之惑人机制所塑造出来的,并因而呈现为一种“虚假的意识”(false consciousness)。所以,若人们更加了解此类机制,或者读过揭露此类机制的理论,他们就可以有道理地把社会共识视作不可靠的东西而抛弃掉。

我们现在的问题是,偏见自身是否可以具有一种功能,这种功能由生成偏见的社会自我平衡机制所赋予。这要取决于社会过程的本质。在进化植入的假设那个例子中——假定有这样的假设——那些假设与事实是充分匹配的,人们因此有优势,能够更容易获得那些事实。权衡信息与理由中出现的社会偏见,是

---

① 因此,我们应该重新解释所罗门·阿施(Solomon Asch)在社会从众问题上的著名试验吗?

由适应重要事实的选择机制所产生的或要服务于这个社会里的广泛目标吗？如果是的话，那么它将支持这种保守主义观点，即既有制度、传统和偏见具有支持它们的强力前设。但我对此表示怀疑。仅当这种选择程序是严格的，并且指导此种选择程序的准则是可欲的时候，我们才能为该种程序所生成的任何偏见赋予合法的权重。① 130

有种保守主义前设，它支持那些长期存在的制度、传统及偏见。除非经过严格的限定，否则从表面上看这种前设是完全讲不通的。奴隶制曾长期存在，女性在社会中屈居从属地位、种族敌视、虐待儿童、近亲通婚、战争、西西里的黑手党等都是如此。某事物长期存在，我们应给予这一事实以多大的分量，这要取决于该事物为什么会持续存在，它所通过的选择检验是何种性质，该检验中又体现了何种准则。（此外，选择所发生的那个环境的性质也有可能发生了相关的变化，因此曾经适应的东西可能不再如此。）既有可能高估一种检验的严格性，也有可能低估它的严格性。马克思主义者认为，自完成推翻封建主义的任务之后，资本主义的社会和制度最近通过的唯一检验便是要服务于统治阶级的利益，因此，这些人决心摧毁掉他们并不理解的东西。（试看马克思主义在人文及文化学术研究中的复兴情况，人们也许会说，马克思主义是反复上演的，第一次是一出悲剧，第二次是一出闹剧。）

即便某物存在的理由是众所周知的，即便该物所履行的有

---

① 接受这样的观点（即所通过的选择程序是严格而适当的），是否本身有可能也代表着（其自身视角中的）由我们的社会提供给社会成员的一种偏见（比照马克思主义对意识形态的看法）呢？这种偏见之所以存在，是因为其反映了某种关于社会生活的持久的、有生命力的真相呢？还是因为其有助于该（种）特定社会之维续、有助于维护该种社会中的支配性群体之优势地位呢？

些功能是有价值的，这依然无法解决“该物继续存在是否是可欲的”这一问题。这个问题取决于我们是否能够设计和建构一种替代物，是否有理由相信这种替代物是更好的，值得冒随之而来的风险。不过，只有当一个社会被迫做出这种或那种改变时，它才会选择进行全面的制度变革。这样一来，社会制定的决策将往往是保守的，尽管这种结果不是把现存之物评价为由那种具有可赞准则的严格选择程序所得来的。

资本主义是经过一系列转变逐渐从封建主义产生出来的，每一个转变都产生了明显的正效益(提高了生产力，扩大了可维持的人口规模等)，并为进一步的转变提供了理由与动力。资本主义是以一种“爬坡”(hill-climbing)方式兴起的。(诸如眼睛这样一种复杂且相互联系的机制是能够进化的，若进化的方式是对每种成分都产生某种改善，即若感光度增加一点都能产生某种好处，即便这些利益小于整个眼睛所提供给我们的全部利益。[1])这样一种路径只能保证产生一种局域最优，但即便资本主义就是如此，若不存在这样的细小步骤，它们一直是改善的，并且最终会导向一种不同的(号称)全域最优的情境的话，那么在面对一种宣称自己所倡导的是全域最优的运动时，资本主义也还是稳定的。如果情况就是这样，那么社会如何达到全域最优就是有问题的了。[2] 当然，我们既聪明又理性；我们高瞻远

131

---

① 参见 Richard Dawkins, *The Blind Watchmaker* (New York: W. W. Norton, 1986), pp. 77－86。

② 不仅如此，Paul David 指出，标准英语打字机的键盘设置在技术上讲是无效率的，但是给定现有的机械资源及人的打字技能的情况之下，若将其改变为另外的一种键盘设置，那么这从经济上来讲是无效率的。因此，如果一种爬坡法果真能导向一种稳定的全域最优的话，这一最优也可能具有它自身的特定缺点，只是与之相互衔接的那种历史性调节使得修正这种缺点是没效率的。参见 Paul David, “Clio and the Economics QWERTY”, *American Economic Review* 75 (1985): 332－337。

瞩,决定跨越某一险滩。但是,达致目的地所需面对的变化越多、越大,则这一目标也就离我们越遥远。那么,为了做这一决定,认为目的地更好的理由得有多强大呢? 当事情在局部状态相当不错时,这种走向未知的航程就充满了太多的重大危险,不管它看来得到了多少理论论据的支持。(因为我们的理论能有多好,我们对社会的理解又能有多好呢?)

因此,毫不奇怪,大胆和彻底的社会"试验"(这包含且需要诸多要素共同进行转变)更容易发生在那些普遍状况令人绝望、几乎没有独立的核心能抵制这类变化的地方。(马克思主义革命发生在俄国,而不是像马克思主义者本来预计的那样,发生在发达的资本主义国家,并且是为经济上欠发达的国家所效仿的。)

我们可以给出两种理由而对一种社会进程抱有信心。首先,这种社会进程是积跬步而展开的,其中的每一小步都能被我们视为有益的,因而也就为后续进展树立了信心。但这也就是说我们尚未达致一种局域最优,因而根本无助于达致一种遥远的全域最优。(亚当·斯密在《国富论》中对资本主义社会所作的理论描述,是在此种进程业已展开且取得了一些积极效果之后。因此,不仅是既往的步骤,而且斯密的理论也鼓励了进一步的步骤。不仅如此,斯密的理论也说服了当局,因为它解释了过去步骤的成功结果——它不"仅仅是理论"。)

采纳一个社会进程的第二种理由是,思路相同的小规模试验(在现存社会里)已获成功。然而,如果一种全域最优只有建立在整个社会层次上——或说在整个国际社会?——才能运作良好,而当体现在一个有着不同特征的社会里就不行时,那么找到前述这种令人鼓舞的局域成功就是不可能的了。即便这些小范围内的实验确实成功了,但仍然存在着这些结果能否外推的

问题。它们能否在大社会中起作用呢？它们能对每一个人而不仅是特选出来的参与者或者监督者起作用吗？

由于自身原因而运作相当好的社会，不会自己独立地达致132一种与局域改善无关的那种全域最优。全域实验不会在那里首先得到尝试。但是，这些实验可以在其他地方试行，并在那里示范其生命力及有效性。而且，通过国际经济的联系及竞争，一个外部范例有助于引导人们对自己的社会做出重大修改。①

除了生物学机制和社会机制塑造我们出于理由而行动外，还有一种我们对自己所做的塑造，即根据我们所选取的目标来修正并指导我们的理由观念和我们对理由的运用——的确，在这一过程中我们运用了生物性地和社会地（以及既往的个人塑造的结果）塑造出的那种能力。无论理由的初始功能为何，我们都可以用我们的能力来运用理由构建理由的新属性，并塑造我们对理由的使用来展示出这些属性。也就是说，我们能够修正和改变理由的功能，由此也改变了合理性的功能。

---

① 有些人主张美国应该以日本的工业关系模式为样板，此观点由于国际经济竞争的结果而得到了人们更为认真的考虑。同样，东欧与前苏联之所以愿意尝试朝着市场资本主义方向进行重大改变，也是源于诸资本主义世界所证实可见的更高的经济繁荣。

5. 工具合理性及其局限

工具合理性够了吗

工具主义者会遇到“对什么足够了呢?”这个问题。他们很高兴我们在提问时似乎已认可了他们的视角。但关键问题在于:工具合理性是否就是**整个**(whole)合理性。

在**因果**决策理论框架内,合理性的工具观可以根据决策理论的术语加以表述,这种框架里的(概率)因果联系观把捉了工具合理性的核心观念:手段-目的关系。但是,我们(以及因果决策理论)必须宽泛地设想这种手段-目的关系,以便把行动与更宽泛意义的行动之间的关系囊括进来,后者是前者所规定的、完成前者的一种**方式**(比如乘飞机去芝加哥是到芝加哥去的一种方法)。我们要实现的目标可能就是履行一个行动,而完成它的方法就是以一种具体的方式去做它;那么,工具合理性(即充分且有效地达致目标)就将包括这一点。因此,工具合理性并不总是包含产生某个**其他的**事物,亦即某种完全不同于行为的东西。工具合理性能够承认,做(或做了)一个行为本身就可以具有某种价值,这种价值(也许做点延伸)就可以归于该行为所产生的价值之中。(当我接下来谈到“因果的”及“工具的”时,我是把做这个行为本身的效用和例示方面也包含在内的。)

工具合理性观是一种强有力且自然的观念。尽管有人对合

理性提出了各种更为宽泛的描述,但每一种号称完整的描述都包含了工具合理性。工具合理性处于所有合理性理论的交集之内(或许再没有其他的东西)。在这个意义上讲,工具合理性就是那种默认理论(default theory),亦即所有讨论合理性的人都认为理所当然的理论,不管他们还有什么其他的看法。我认为其中还有其他的东西。合理性的工具论看来并不需要证成,然而其他的每种理论都需要。其他每种理论都必须提出理由,证成其所划定的东西确实是合理性。工具合理性就是那种基础状态。而问题在于它是否就是整个合理性。

134 有人会反驳说,对工具合理性的任何拓展都是无法被证成的,而且拿那些本身不是纯粹工具性的程序来证成这种拓展,这就回避了问题的实质。(如果非工具性程序是由工具性程序来证成的,那么,它们不就总是具有起源上的工具性吗?)因此让我们追问:为什么我们应当在**工具意义上**是理性的。为什么人们应以最有效率和最有效的方式来追求他们的欲望或目标呢?因为这是最可能以最小的代价实现其目标或满足其欲望的方法,因此能够获得总体上目标和欲望的最大实现。但他们为什么应当追求他们的目标或满足他们的欲望呢?因为这些东西是他们欲望做的事情。但是,他们又为什么应当满足**那个**欲望呢?存在任何非循环的答案,即一个不以结论作为前提而证成工具合理性的答案吗?

如果说其他合理性模式无法以非循环的方式证成自身的话,那么工具合理性也是如此。[①] 所以,循环论证本身并非是对

① 支持工具合理性的观点大概也会遵守那些能有效地达致认知目标的标准。然而,彻底的工具主义者不会认为,接受或追求那些特定的目标是一种合理的要求;因此,如果他是对的,那么他的观点**至多**只能(合理地)说服那些碰巧与其共享这些目标的人。

于其他合理性模式的一种结论性批评。无论如何,工具合理性基于什么东西拒绝循环性是不清楚的。是否有经验证据能够表明:当我们在实际上被迫运用循环证成的情境中而使用它时,它会导致在实现欲望(或获得真理)上比其他(特定)程序做得更差吗?

愿望为什么应当被满足呢?我提出这一问题并不是为了宣称工具合理性(原则上)不能是合理性的惟一内容。工具主义者可以把满足愿望视为一个给定的目标,而不是一个合理的目标,那么,若批评者要求把满足愿望本身表明为合理的,这就是引入了一条工具主义者不会接受的标准。我提出工具合理性的证成这个问题,也不是为了质疑工具合理性的合法性。(我对此毫无怀疑,至少在指向合理目标的工具合理性情形中是如此。)但是,我还是要宣称,还存在着其他的合理性模式,因此合理性概念并未被工具性概念所穷尽。我之所以提出工具合理性的证成问题,是为了回避早前工具主义者的这种指责:其他合理性模式的证成必定会陷入循环。他们在这一点上的处境是相同的。

对合理性的一种因果工具论说而言,我们的合理性标准必须依赖于我们(具有我们的能力、力量、缺陷及弱点)对这个世界特征的看法和对我们是何种人的看法。(就更为宽泛的而不仅仅是工具性的论说而言,仍然会存在一种类似的但局部的依赖关系。)无疑,如果合理性是为效果所确定的,那么什么东西实际上(或一般地)将具有某种效果就是一个经验性问题。涉及世界中的一类人时,有些标准就是能(因果地)达致一个目标的最有效标准;而在另外一种世界中,会有另一些标准是最有效的。显然,这里有种相互作用。我们使用某一时刻的标准去发现世界与我们的特征,继而再基于这种新理解去修正或改变我们的标准,以使这些标准(当我们使用时)能在(我们新理解的)这个世

135

界中(最可能)最为有效。这个过程会持续不断,因为这些新标准还会引起我们对这个世界和自己的看法上继续做出修改,而这又会得到更新的标准……依此类推。所以说,我们对世界和自己的看法,还有我们对于何为合理性的观念,都是处在不断的相互作用之中的。

我们也许会把一种**纯粹的**合理性理论看作是任何类型的存在物在任何类型的世界中(也就是每种存在物和每种世界中)都应该遵守何种标准的理论。但是,即使*存在*这样的一种标准,它们也不大可能有多少内容,或者能带我们走多远。它们可能只是为我们的谋划提供一个起点而已,亦即,在一个世界内,我们拿这些纯粹的标准去构建一种世界观与自我观,尔后据此修正我们原来的标准,在这些修正后的标准基础上再去构建一种新的世界观与自我观,依此类推。然而,没有任何理由认为,在我们具有内容更丰富的标准的实际历史中,背后曾经有过这种标准。我们的标准天生就是不纯净的,试图把我们从这种原罪中救赎出去的哲学努力,迄今为止均是以失败而告终的。

还有一种相互作用值得一提。我们的决策原则与推理原则是交织在一起的。我们会推理应遵循哪些决策原则——本书第二章即为一例。我们也可以决定应当遵循哪些推理原则。决定遵守一组具体的推理原则的策略是一个行动方案。因而,对于涉及不同组推理原则的两种行动方案而言,我们便可以通过一种决策原则的评价而决定何种方案更好。这样一来,何种推理原则会被认为是更好的,这又将取决于采用何种决策原则。

这一点明显适用于这样的人,他们想用各种推理原则得到真理的可靠性来证成它们,亦即最可靠的推理模式便具有最好的证成。我们还想考虑,当推理原则**出错**时,亦即此种推理原则未能得到真理时,我们的境况会如何。如果有一种原则最为可

136

靠,但出错时会有灾难性后果;而另一种原则虽然不那么可靠,但出错时也不会有如此糟糕的后果。那么我们最好还是选择后者,为了其他的利益而牺牲一些可靠性。而且,由于智力(intellectual)原因或个人性原因,有些类型的真理比其他类型的真理对我们更有价值,那么,我们可能就会偏爱一种能很好地达致前面*那些真理*的方法,即便这一方法的总体可靠性比不上另外一种。因此,一种推理方法本身可能是不仅仅要受到达到真理的概率影响,还要受到决策理论中的期望效用考虑的影响。还有其他方面的考虑可能也会介入,由此可能会使得简单的最大化期望效用不适合来选取推理原则;那么这就会使用另外一种不同的决策原则。

一种决策原则是通过一种推理方法(暂时性地)确立起来的,一种推理方法又是通过一种决策原则确立起来的。探究决策原则与推理原则所有可能的搭配,以搞清楚在何种情况下有哪些搭配是相互支撑的,这个谋划野心太大,没有可能性。但我们还是应当希望,在自己的境况下,在我们最为偏爱的任何组决策原则与我们认定最有说服力的任何组推理原则之间,存在一种显著的相互支持关系。而且,此二者间的错位就应当成为变革的契机。我们欲求的是一种历时性的收敛,我们能合理地称之为自我纠错。

即便把合理性仅仅作为工具合理性加以理解和解释,这种合理性也可以在部分程度上是本身就有价值的——参见杜威(John Dewey)对手段成为目的的论述——并因而具有*内在*价值(intrinsic value)。可以设想,这种合理性的*本质*将完全是工具性的,但其*价值*却并非如此。我们重视人们以回应得胜理由的那种方式来合理地相信和行动,我们认为此种方式本身就是好的和值得推崇的——这也许是因为此种决策和相信的方式运

用了我们高超且复杂的能力,并且表现出了这种能力;也有可能是因为此种方式体现了一种值得推崇的、有原则的整体性,它不是以瞬时的直觉或欲念作为指引,而是根据理由来指导行动和信念的。不仅如此,海德格尔强调过这种论点,常常得到充分利用的工具性器具能够成为我们自身的拓展;当我们与世界互动时,我们的边界(boundaries)可以通过这些器具而拓展至它们的目的。因此,原本彻头彻尾的工具合理性和(我们第一章论述的主题)原则,在被充分利用后就变成了我们自己的一种延展,也作为我们的身份(identity)与所是(being)的重要组成部分而被吸纳进我们自身。

我们还可以把合理性视为**理解**(understanding)的路径,不仅仅是作为理解的一种手段,而且也是理解本身的一个构成或组成部分。要理解有些复杂体系,(至少对我们来说)只能通过一种清晰的理论,我们能够知道这种理论与其他理论间的关联,知道支持它的那些理由,我们还可以追踪到它抵御反驳的能力(以及它与反对观点相遭遇的边界)。请注意,合理性不仅仅是寻找这种理论的一个工具,而且它本身也是"理解这种理论是什么和这种理论所描述现象中"的定义性成分。现在若理解是我们在部分程度上因其自身的缘故而重视的——不管我们开始是为什么重视它的——且合理性要作为一种成分进入这种理解的本质的话,那么,这种合理性也可以在部分程度上因其自身的缘故而被我们所重视。

137

然而,若构建的决策理论除了容纳因果性期望效用之外,还容纳象征性期望效用和证据性期望效用,那么我们的合理性观念就已经得到了拓宽,超越了单纯的工具性。合理性不只是一个(很可能)导致或产生什么东西的问题。我们在第二章得出,因果决策理论仅凭自身无法成为一种完全恰当的合理决策理

论。我们讲过,一种合理决策会最大化一个行为的决策价值,而决策价值是因果效用、证据效用和象征效用的加权和。然而,工具合理性却完全为因果性期望效用观念所把捉和穷尽。既然因果性期望效用只是合理性的一个方面,那么这也就是说,除工具性以外,合理性一定还包含其他的因素。当然,人们一直都是如此认为的。在人类历史上,证据性因素和象征因素一直都能产生十分重要的社会后果(再次回想讨论"得选标志的加尔文主义观点在资本主义发展上怎样起作用"的那些文献)。

在合理性信念(belief of rationality)的两阶段理论中,我们已经把这些非工具性的因素囊括进来了。第一阶段是把那种可信值不如相竞争陈述高的陈述排除在候选之外。这些可信值本身是由一种符合(概率式)事实关系的网状连结所决定的,其权重转换在获得真信念或其他认知价值上是完全工具性的。但是,第二阶段是要决定,我们是否要相信可信值不为相竞争陈述所胜过的那些陈述,而这显然并非工具性的。我们此时要衡量的正是相信该种陈述的决策价值,亦即,这种陈述的证据性、象征性和因果性效用的加权和。(工具主义者会把规则 2 而非规则 5 用作是因果决策理论的表述。)

我一直假定,若不仅仅涉及因果性期望效用,而且还涉及决策价值,那么计算就不完全是工具性的。因为这样一来,我们关注的就不仅仅是产生的结果,即关于产生的概率,还有所标示和象征的东西。然而,工具主义者还能够宣称,这也只是单纯的工具性计算吗?如果我们设定的目标是最大化决策价值,设定与此相关的效用函数是线性的,那么行为人不是会选择那个因果上最有效地产生最大决策价值的行动,因此再一次是工具性地行动吗?这里看来相关的概率并不全都是因果概率;行动与其象征效用和证据效用之间的关系也不是因果性的。然而,我们

138

对工具合理性已经做过拓宽描述，包含了行为是做另一行为的一种方式的那种方式。那么，为什么不能把一个行为与证据效用还有象征效用之间的那些关系也纳入到工具性中呢？这也许会使工具性问题变得无足重轻，但一个有趣的观点仍然成立，尽管现在需要重新表述。如今有些行为的目标变成了追求决策价值的最大化，这不仅是一种非工具性的目标——我们早就知道工具性行为的目标本身可以是非工具性的——而且这个目标所描述和命令的行动与目标处于非工具性关系中。由此，只有通过把一个工具性行动的目标视作该行动与其他目标之间处于一种非工具性的关系中，这一更宽泛的概念才能“拯救”工具性。但就我们的目的而言，这正是构成非工具性的因素。

当陈述的可信值是在处理正反理由的网络之中来决定时，其可信值又会如何呢？若在网络中得到的权重只是由这些权重在因果地获得各种认知目标(如真信念、解释力和简明性)会多么有效所决定的，那么就可因此说可信值是完全工具性的吗？这取决于反馈规则(亦即确定可信值的那种处理系统网络中的习得规则)的性质。反馈回来的内容是什么，它是根据何种修改规则来这样做的呢？证据因素和象征因素在此也有存在空间吗？

工具合理性并不是合理性的全部，我对此的论证一直都不是冷漠的(disinterested)。如果人类只是休谟意义上的存在者(Humean beings)，那似乎是在贬低我们的地位。人类是惟一不甘于只当动物的动物。(既然我这个观点是有出发点的，因此你——也包括我自己——应当注意，要纠正此观点在处理各种理由时所附带着的任何偏见。)因此，我们的行为并非全部旨在满足我们既定的欲望，这对我们具有象征重要性。我们已经看到，原则为我们提供了一种控制及重塑我们欲望的手段。(尽管

如此，当康德使原则与欲望脱离关系，指望仅仅尊重原则本身就能产生行为的时候，他还是对原则要求过多了。）

139

我们不只在工具意义上是理性的，表现的方式就是除了关注所引起或产生的东西之外，还关注各种象征意义。工具合理性的倡导者并不能轻易地声称，因为他不具有任何相关的合理性准则，所以这种关注是不合理的——那么为什么这种关注就比其他的关注更加不合理呢？象征意义是一种超越普通欲望之因果连结的方式，我们这样做对我们具有象征重要性。此时，工具主义者会笑容可掬地问我们：这是否意味着，依据象征意义而行事，亦即考虑象征效用就是一种超越休谟式连结（Humean nexus）的手段呢？或许是的，但即便这种考虑不能完成这种超越，它也能象征这种超越。① 那么，即使对于我们的信念形成与维持过程来说，我们也可以不仅考虑这些程序因果性地产生了什么，还可以考虑它们同时象征着什么。我们在第一章中对于原则的讨论主要是工具性的；我们讨论了原则所能发挥的各种功能。这里我们看到了一种可能的元功能（meta-function）——亦即它超越了对其他功能的服务——因此，遵守原则也可以带有一种象征效用。②

---

① 既然朝向象征意义的这种导向本身对于我们而言便具有一种象征意义，那么放开象征效用的实际效果不谈，象征效用也能够支持其自身。遵守要考虑象征效用这一原则的行为本身就具有一种象征效用，因此可以归为该原则中的一个范例。工具合理性也可以自我支持和包摄其自身。关于自我包摄（self-subsumption）的问题，参见我的 *Philosophical Explanations* （Cambridge, Mass.: Harvard Univ. Press, 1981）, pp. 119－121, 131－140。

② 请注意，我们在网状结构之内已经有了两套机制：应用证据性条件概率的标准贝叶斯公式；还有应用因果概率的因果版本。有没有哪种途径能把这两套机制以及象征考虑都联结起来呢？作为一个粗略的近似机制——实际处理网络要用更为复杂的描述——我们可以把依据 e 而相信 h 的可信度，即 $cred(h, e)$ 视作三者（即因果性的贝叶斯比率［附有假设条件］、标准的贝叶斯比率［附有条件概率］（转下页）

我所主张的是，除了工具合理性以外，还有其他合法的合理性模式——此中包括证据性和象征性的合理性模式——但这里仍然存在一个问题，在不同的合法模式中哪种优先。还有我们将选择确立何种优先性的问题。

合 理 偏 好

工具合理性观念宣称穷尽了合理性之全部领域，对此当然存在着常见的批评。当某事能因果有效地实现或者满足各种给定的目标、目的、愿望和效用时，它对于后者而言就是工具地合理的。然而，工具合理性观念却根本没有任何办法让我们去评价那些目标、目的和欲望本身，除了这些目标与欲望是否也是工具性有效地获得进一步的给定目标之外。即使对于相信真理这样的认知目标而言，我们似乎也只有一个工具性证成。对于目标与欲望的实质合理性而言，我们目前没有任何恰当的理论来消解休谟的这个表述："只擦伤一下指头就可拯救整个世界也不干，这也与理性并行不悖。"①

140

我想采取几个暂时的步骤，以得到一种目标与欲望的实质

（接上页）和象征成分 $sym(h, e)$）的加权和。（然而，不清楚的是，这里恰当的象征成分指的是什么［亦即对以 e 为基础相信 h 象征着什么］。它象征着能相信的真理程度吗？）这些权重正是我们在决策理论中（即在我们的 DV 公式中）所使用的那些权重吗？（如果是这样）那么，个人就可以用**这些**可信值去排除那些不值得相信的陈述，即其可信值比与之不相容陈述的可信值更低的那些陈述。此后，他将使用进一步的规则来决定要相信可允许陈述中的哪一个。

① David Hume, *A Treatise of Human Nature*, ed. L. A. Selby-Bigge (1888; Oxford: Oxford Univ. Press, 1958), bk. II, pt. III, sec. III, p. 416. 休谟继续道："为防止一个印第安人或一个陌生人不受一丁点不适，我选择完全牺牲自己，这也完全不悖于理性。我偏好明知是更少的利益，这一点也不悖于理性。"

合理性的理论。我要强调一下,我的目的并不在于要采纳我将要提出的那些具体条件,或者是捍卫它们的具体细节。毋宁说,我希望表明,我所讨论的这类条件有着什么样的成功希望,我们可以从何种方向来超越休谟。

一如既往,作为一个哲学家,我从对形式的观察来开始对"内容"的讨论。我认为,超越休谟一小步的东西,这也是休谟不需要反对的,是各种偏好如何结合在一起的各种约束。这些约束表述在决策理论中所提出的标准冯·诺伊曼-摩根斯坦条件或其各种变体之中。比如说,偏好是传递性的,亦即当两个选项有两个相同的可能后果,只是每个后果赋予的概率不同时,那么赋予更可取后果以更高概率的那个选项更可取。[①] 有些这样的条件是工具性地证成的,比如说偏好是可传递的那种"吸钱机[money pump]"论证;[②]而其他条件的提出则是因为其表面上的规范吸引力。(除非后面这些条件也能给予一种工具性证成,否则这不就是超越工具合理性的一步了吗?)当代的决策理论将

---

① 参见 John Von Neumann and Oscar Morgenstern, *Theory of Games and Economic Behavior*, 3d. ed. (Princeton Univ. Press, 1953), 附录;R. D. Luce and Howard Raiffa, *Games and Decisions* (New York: John Wiley, 1957), pp. 12-38。

② 这个思想是,假如个人的偏好不具有传递性,例如他偏好 x 于 y,偏好 y 于 z,也偏好 z 于 x,那么,当他以 z 为起点时,他就会花费一小笔钱而使自己的状态改进到他更喜欢的 y,接着会再花一小笔钱把状态改进到相对于 y 而言更喜欢的 x,然后会再花一小笔改进到相对于 x 而言更喜欢的 z——这正是他的起点 z。因此,最后的结果就是他有净损失。参见 Donald Davidson, J. McKinsey and Patrick Suppes, "Outlines of a Formal Theory of Value", *Philosophy of Science* 22 (1955): 140-160, 他们将此论证归于 Norman Dalkey。这个论证假定个人总是乐于依凭每一种孤立地考虑的个别偏好而行事,无论他是否知道所有这些偏好组合到一块的情况,亦即无论他是否预见他的系列行为将导致他进入这类困境中,他还是会一直愿意**反复地**依凭每个个别偏好而行事。这显然是一种讲不通的假设。鉴于吸钱机论证是为了证成偏好应该是可传递的这一规范条件,那么,确切地建构出**这种论证**所依凭和预设的那种规范性条件,将是一种有益的尝试。

此作为超越休谟的一步：虽然它没有说任何个别的偏好是不合理的，但却可以说放在一起的个人偏好集是不合理的。让我们假定存在这样的一些规范原则，它们规定了把几个偏好放在一起的结构，是合理性的条件。（对有些冯·诺伊曼-摩根斯坦条141件，在文献中有一些公认的反例和反驳；因此这里的观点不在于使用那些特定的条件，而是使用某些恰当的条件集。①）

I. 对于偏好及偏好与概率的关系而言，行为人满足冯·诺伊曼-摩根斯坦或其他恰当规定的条件集。

这表明，个人的偏好要是合理的话，至少还必须满足一个进一步的条件，亦即，她必须偏好满足那些规范性条件而不是不满足它们。实际上，无论是偏好、行为还是信念中的合理性，对合理性的任何有效结构条件 C：

II. 行为人偏好满足而不是不满足合理性条件 C。②

[此条件应陈述为初定(prima facie)条件或者附上一个进一步的条款，正像下面许多条件也都应该这样做的那样。假设个人一直满足无差异是可传递的这一条件，或不相信"可信度比不相容陈述低的任何一个陈述"这一条件，那么他就会被杀掉。一旦他知道这一点，那么他很可能偏好不去这样做]我们可以假定，既然这个人在工具意义上是理性的：

① 此外，我们还可以试着描述偏好(单个或一起)形成与修正的一个(规范性)程序，并且检验这一程序(如果无限期地执行的话)能否产生一种冯·诺伊曼-摩根斯坦效用函数。由此，冯·诺伊曼-摩根斯坦效用理论就可以被视为对于特定程序之结果的目的状态描述，至少在极值上是如此。检验的结果可能是，我们没有什么特殊的理由要把一个程序贯彻至极限。

② 我们是否应该区分由合理性所要求的两种欲望观念呢？第一，可合理地拥有的欲望；第二，如果合理性自身是可欲的或有价值的，可合理地拥有的欲望。如果忽视了后者，则我们就无法弄清楚，为什么"拥有那些可合理地拥有的欲望"本身是可欲的。我们是否可以这样说呢？

III. 其他所有情况相同,行为人会欲求能够满足合理性条件 C 的那些手段和先决条件。

这些合理性条件 C 不仅关注各种偏好的结构,而且关注信念合理性的任何恰当的结构性条件。因此行为人会欲求合理信念的手段与先决条件,亦即她会欲求能有效地分配可信值(和决定持有一个具体信念的效用)的手段与先决条件。

当个人偏好选项 x 于选项 y 时,又偏好没有这种偏好,也就是说,他也偏好不偏好 x 于 y 而不是偏好 x 于 y,那么,此人便欠缺一种合理的完整性。当这样一种二阶偏好与一种一阶偏好相冲突时,应当改变何种偏好则是开放的。清楚的是,这些偏好没有很好地结合在一起,而且理性的人会偏好情形不(继续)是这个样子的。① 我们由此有一个条件,行为人要具有一个特定的三阶偏好,也就是偏好不要出现偏好的这种冲突。设 S 代表这种冲突情境,那里行为者偏好 x 于 y,但同时却偏好不持有这个偏好。也就是,设 S 代表: $xPy\&[(非(xPy))P(xPy)]$, 那么

IV. 其他一切情况相同,对于每个 x 与 y,行为人偏好非 S 而不是 S。 142

这并不意味着个人无论如何都必定选择非 S 而不是 S。一个瘾君子欲望自己不要欲望海洛因,但又可能知道自己无法克服欲望海洛因的这个一阶欲望,因而对这种冲突的惟一解决之道只能是放弃自己的二阶欲望,即不想要那个一阶欲望的欲望。

---

① 对二阶偏好的讨论,参见 Harry Franfurt, "Freedom of the Will and the Concept of a Person", *Journal of Philosophy* 68 (1971): 5 – 20; Amartya Sen, "Choice, Orderings and Morality", 重刊于他的 *Choice, Welfare, and Measurement* (Oxford: Basil Blackwell, 1982), pp. 74 – 83; Richard Jeffrey, "Preferences among Preferences", *Journal of Philosophy* 71 (1974): 377 – 391。也参见 Gilbert Harman, "Desired Desires", 载于 *Value, Welfare, and Morality*, ed., R. Frey and C. Morris (即出)。

尽管如此,他仍有可能倾向于保留这一欲望的冲突,因为有了它,他将不会那么彻底满足毒瘾,或他的毒瘾没有那么糟糕。①

休谟宣称所有的偏好都是同样合理的。但是,偏好是什么和偏好是为了什么,对它们的理解也许会使进一步的条件成为恰当的。晚近的各种理论已经把偏好理解为选择某一事物而非另一事物的一种倾向(disposition)。② 各种偏好的功能,亦即进化把对它们的能力植入我们的理由,最终是为了得出偏好选择。但人们只有在某些条件下才可以做出偏好选择:活着、有能力知道各个备选项、有能力做出选择、有能力完成的选项的行动、不面临任何使得前述能力不可能施展的各种干涉。这些便是偏好选择的先决条件(手段)。现在,人们并非必须偏好这些条件的继续存在;比如说,有些人有理由想死。但我认为,这些人需要有一个理由;因为根本没任何理由而单纯想死是不合理的。这里有一个前设:即要做出任何偏好选择的话,人们就得偏好于满足那些偏好选择的必要条件;她不一定实际上持有这一偏好,但对此她需要理由。

V. 在没有任何特定反对理由的情况下,行为人偏好于满足使她的任何偏好选择得以可能的每一个先决条件(手段)。

因此,人们偏好活着而不是死亡,偏好有知道其他备选项的能力而不是失掉这种能力,偏好有实施选择的能力而不是让这种能力被毁坏等。③ 再次,我们可以加上

---

① William Talbott 和 Amartya Sen 分别向我指出了这一点。

② 我对描述这种划定的诸多困难的讨论,见 Robert Nozick, *The Normative Theory of Individual Choice* (1963; rpt. New York: Garland Press, 1990), pp. 39 - 48, 70 - 78。

③ 把这一条件构建为这样的一个前设,即人们在没有不支持它的理由时就成立,这可以避免一种反对意见,这种反对意见是针对我在"On the Randian Argument"(*The Personalist* 52, no. 2 [1971]: 285 - 286)一文中所提出的更强原则的。

VI. 其他一切情况相同，行为人偏好使作为偏好选择之先决条件的那些能力不被这样一种惩罚（一个极不可取的选项）所干涉，这种惩罚会使他在其他情境下偏好永远不去施展这些能力。

我认为，对于理由还有更多的东西可谈。（我只是临时性地提出这点；解决这个问题还要做更多的工作。）假定我毫无理由地偏好 x 于 y。那么为了获得我所偏好的某个其他的东西，我会愿意，也应当愿意，颠倒我的这个偏好。假使在我的能力范围内，为了得到 25 美分，我应当会愿意颠倒我的偏好，即从现在开始偏好 y 于 x。然后我就会从偏好 x 于 y 的这个情境，转变为偏好 y 于 x，并且得到那 25 美分的这个情境。难道我不会偏好后一种情境于前一种情境吗？[①]可能不会，可能我**强烈地**偏好 x 于 y，且根本没有任何理由。在我看来，没任何理由的强烈偏好是一种反常。既然我具有这种偏好，那么我将按它去做；但若我没有任何理由持有这种偏好，却为了追求它或者维持它而付出代价，那么这种坚持就是不合理的。或者，假设我偏好“偏好 x 于 y”而不是没有这种偏好，且我的偏好足够强可以胜过 25 美分。这样这个对偏好 x 于 y 的二阶偏好便使我不愿放弃该偏好。但是我为什么要持有这个二阶偏好呢？我想说的是，与任何任意的一阶偏好不同，二阶偏好的背后是需要理由的。除非个人有某种理由来偏好 x 于 y，否则他对“偏好 x 于 y”的这个二阶偏好便是不合理的。也就是说，他必须有理由来偏好具有那个一阶偏好——或许是他妈妈告诉他的，或者该偏好现在成

143

① 让我们暂且不考虑这种情况：假如另一个人给了我 25 美分，我可能偏好自己的偏好不受那种外在资源所左右。

了他的身份的一部分，因而是他不想改变的东西[①]——或者有一种直接的理由偏好 x 于 y，亦即有关 x 与 y 之特性的理由。但什么是直接的理由呢？这个语境下的理由必须是不同于偏好的东西吗？它至少必须是能起到理由作用的一种偏好，亦即，是**一般性的**(general)偏好，尽管是可挫败的偏好。我们一般认为，有个理由来支持偏好 x 于 y 就会涉及对知道 x 的某个特征 F，以致，一般而言，其他一切情况相同，在 x 所属的那类型的事物中，你会偏好含有 F 的事物而不是那些不含有 F 的事物。[②]（偏好冷饮胜于热饮，并不要求偏好寒冷的房子而不是暖和的房子）

VII. 如果个人偏好 x 于 y，或者(a) 他为了一丁点好处愿意转变为偏好 y 于 x，或者(b) 他有理由偏好 x 于 y，或者(c) 他有某种理由去偏好偏好 x 于 y 而不是没有那种偏好。

我不是说个人的*所有*偏好都要有理由——不清楚对于最顶端偏好要说些什么；或许它们是由其下各种偏好所固定的——但是当人们不想改变他们的一阶偏好时，那么它们确实是需要理由的。一旦我们着陆在对各种偏好的理由领域之内时，我们就能考虑，那些更普遍的理由与不那么普遍的理由之间是如何相关的，或者为各种理由施加一致性条件，等等。这样一来，我们就打开了一条大道，可以为偏好加上更进一步的规范性条件，至少对人们不愿轻易改变的那些偏好而言是这样的。尤其是有偏好反对"前面提到的偏好选择的先决条件"的情形下，个人需要的不仅仅是理由，还得是有一定分量的理由。这就至少意味着，这些理由必须与此人的众多其他偏好交织在一起，甚至是在

144

---

① 我把这个关于同一性的观点归于 Howard Sobel。

② 当它是 y 的一个最重要的负面特征的时候，读者可以提供修正。

不同层次上交织在一起。①

既然得悉偏好与欲望的起因不会导致你不再(想要)具有它们,我们知道此点后,可能还想加上愿望与偏好是处于均衡的这一条件。愿望与偏好经得起知道其起因。②

VIII. 个人的愿望与偏好(在知道其起因后)处于均衡状态。

既然偏好与欲望是要实现或满足的,若个人的偏好具有这样一种结构,即总是这山望着那山高(有 x 时想要 y,而有 y 时便想要 x),那么他注定是要失望的,更多的是,老婆并不总是他人的好。因而

IX. 不存在任何这样的 x 与 y,个人在情况 y 时总是偏好 x,而在情况 x 时又偏好 y。(他的条件偏好不是这样的:即对于某个 x 与 y 来说,给定 y 时,他偏好 x 于 y;当给定 x 时,他则偏好 y 于 x。)

欲望并不就是偏好。从偏好上升为欲望的步骤中存在着一个过滤或者处理层次——正如(我们将要看到的)从愿望上升为目标有着另一个层次。我们可以说,合理的欲望就是那些有可

---

① 不合理欲望或信念的一个标识大概就是,它没有经受全局性的控制与修正。它只是独自坚持,拒绝与其他欲望与信念进行整合。参见 David Sharpiro, *Neurotic Styles* (New York: Basic Books, 1965)。我猜想"催眠后暗示"(posthypnotic suggestion)也是同样的情况,它不能融入到整体的信念与欲望网络中并在此网络中得到修正。

② 参见我的 *Philosophical Explanations* (Cambridge, Mass.: Harvard Univ. Press, 1981), pp. 348 - 352, 714 - 716。有些论者构建了进一步的条件来把偏好、愿望与知识关联起来;这些人不仅关注知道偏好与欲望的原因,还关注知道它们的结果,以及它们与其他一切事物的相互关系。例如,参见 Richard Brandt, *A Theory of the Good and the Right* (Oxford: Oxford Univ. Press, 1979), pp. 110 - 129, 149 - 162; 批评方面则参见 Allan Gibbard, *Wise Choices, Apt Feelings* (Oxford: Oxford Univ. Press, 1990), pp. 18 - 22。

能实现的欲望，或者至少是那些你认为有可能实现的愿望。让我们十分审慎地说

X. 个人不会持有她明知不可能实现的欲望。

个人偏好不借助器械飞翔也许是完全正确的，但假如他**欲望**此事，那便是不理性的。[尽管希冀(wish)这是可能的，这也许是合理的。]不像单纯的偏好那样，愿望会进入某种决策程序。它们必须通过某些不是孤立的可行性检验：你的诸欲望必须有共同实现的可能性。一旦发现它们是不可能共同实现的，欲望必须加以改变；尽管一个修改了或放弃了的愿望仍可作为一种偏好。①

145

依次来说，目标既不同于偏好，也不同于欲望。② 持有或接受各种目标，就是为了用它们在选择情境中过滤掉不能很好地或根本就不能为它们服务的那些行动。作为一种能力有限的存在物，我们不能在每时每刻思考与评价每一种可选择的行为——你现在可以试着列出一个可选行为的清单——因此，这样一种过滤装置就是至关重要的。不仅如此，我们还可以利用目标来产生要严肃考虑的行动，亦即那些**确实**服务于目标的行动。③ 目标为行为的结果提供了显眼的维度，这些目标在评估

---

① 我并不是关注**欲望**这个词，或者是关注这个词实际上应用的那些现象。或许正是**目标**这一术语标明了那些我们必须不能是明知其无法被共同实现的东西。真正重要的是在于越来越多的约束，而不是标签之间的概念区分。

② 各种目标及其功能产生了诸多问题，对之极有启发性的讨论，参见 Michael Bratma, *Intention, Plans, and Practical Reason* (Cambridge, Mass.: Harvard Univ. Press, 1987)。Bratman 在意向(intentions)主题下讨论了许多问题。

③ 参见 Helmut Jungermann, Ingrid von Ulardt, and Lutz Hausmann, "The Role of the Goal for Generating Actions", 载于 *Analysing and Aiding Decision Processes*, ed. P. Humphreys, O. Svenson, and A. Vari (Armsterdam: North Holland, 1983), 特别是 pp. 223 - 228。

结果效用时将很有分量。既然目标具有多重重要的功能,那么人们就会期待,对一个长期稳定的重要目标而言,我们会把为数不多的敏感渠道之一用于它,关注有望实现它的各种路径,监管我们目前的进展,等等。①

目标是怎样产生的呢?它们是如何遴选出来的呢?它们是从偏好、愿望和有关概率(probability)、可能性(possibility)、可行性的各种信念所组成的矩阵中得出来,这种想法看来是讲得通的。(然后目标会重组我们的愿望与偏好,更看重某些欲望与偏好,也会颠倒其中的一些偏好,因为颠倒的偏好会契合或促进该目标。②)一种可能性是目标是在期望效用理论的应用中出现的。对每一个目标 G_i 的追求,我们都会将之处理为对各种结果具有自己的概率分布的一个行动,并计算出这个"行动"的期望效用。我们会采用具有最大期望效用的那个目标,然后用它来产生各种可选项,排除其他的选项,等等。

对这种把各种目标契合进一个期望效用框架的简单做法,有着一种反驳。一旦把 G_i 树为目标,就将产生很大的影响。G_i 作为一个目标起着排除装置的作用,其地位已截然不同于只在效用上略小但很接近的选项 G_j。③ 一个原本细微的差别现在产

---

① 参见我的 *The Examined Life* (New York: Simon and Schuster, 1989), pp. 40 - 42。

② 是否还存在着进一步的因素,它与目标的关系,就像目标与欲望、欲望与偏好之间的那种关系,还涉及另外一个层次上的处理与过滤吗?

③ 回忆 Isaac Levi 把信念处理为严肃可能性(serious possibility)的一个标准,因此个人无需去考虑该信念有出错的可能,或者考虑其出错的情境。(见上面 p. 96)。Levi 的相信某事的规则可以依凭一个极为微小的差别判定此点;不过,一旦某物被弄成了一种信念,它就将产生重大的影响。然而如果我们返回到它被弄成"信念"之前的那个情境,那么它与另一个假说,即另一个可能的信念之间的差别将一直是非常小的,显然不够直接在效果上产生如此截然不同的差别。

生出了重大差别。重大的差别(例如由某一事物设定框架,而其他东西被排除在外)看来应当立基于既有的重大差别。[①] 我们来考虑由蒙哥马利(Henry Montgomery)提出的一种描述性决策理论(descriptive theory of decision)。在这一理论中,个人通过找到一种占优结构来证成一种选择。为了使一个行动在所有(得到考虑的)特性上弱占优于所有其他行动,她利用了诸如特征组合与修正、选项蜕变等机制。由此,因为有一个行动显然是最好的,所以冲突便得以避免;她没有任何理由去选择其他的行为。[②] 但这种占优总是能够在行动之间设立起一道足够重要的**鸿沟**,做出具有重大效果的定性差别,从而适用于目标的形成吗?然而假若两个行为总共只有六个维度,即使它们在五个维度上都能持平,(仅)在第六个维度上有一个行为略微胜出,该行为就能够弱占优于另一个行为。因此,即便在这种框架内,我们看来需要的也不只是弱占优;我们需要的也许是在一个维度上**大大**胜出或者在许多维度上都胜出。

146

我们再次回到期望效用框架,之所以选择目标 G_i,可以说

---

① 至少当没有偏离这一规范的实践必要性时,诸如,在一个名额有限的项目中,那最后一个被接纳的人与那头一个被拒绝的人之间的差别就应该是如此;第二,没有必要把许多只是稍有不同的实体放入名额不够的分类范畴中去时。

② 参见 Henry Montgomery, "Decision Rules and the Search for a Dominance Structure", 载于 *Analysing and Aiding Decision Processes*, ed. Humphreys, Svenson, and Vari, pp. 343 - 369, and "From Cognition to Action", 载于 *Process and Structure in Human Decision Making*, ed. Henry Montgomery and Ola Svenson (New York: John Wiley, 1989), pp. 23 - 49。Montgomery 认为期望效用最大化规则是另外一个问题,因为它把所有信息都纳入了考虑。但是请注意,这个公式是一种把信息组合(合并?)进一种特质(即**期望效用**)中的一种方法,并据此说某一行为在所有相关特征上都胜于或者占优于另一个行为,因为如今只有这样的一个特质(亦即期望效用)了,而且(在那个层次上)不再有理由来反对那一效用最大化行为,或有理由支持另一种行为。

不仅仅是因为它拥有最大的期望效用，而且是因为它**决定性地**打败了其他的备选目标。对于任何 j，都有 $(EU(G_i) - EU(G_j))$ 大于或等于所规定的某个固定的正值 q。(但是，这里还有一个类似但较小的问题，即 G_i 决定性地打败了其他目标，但在是否决定性地打败这点上却没有任何决定性差别；$(EU(G_i) - EU(G_j))$ 的各个值之间的差别或许刚刚达到或没有达到 q。

在某种程度上说，把某物确立为一个目标就是采纳这样的欲望，即要找到从现状实现该目标的一条可行路径。[①] 因此

XI. 若个人明知不存在任何可行的路径(无论多长)可以从现状去获得一个目标，他不会持有该目标。

不仅如此，我们可以说，一个理性的人不仅仅只是具有偏好与欲望，而是会具有能找到可行路径的目标。她将过滤掉那些不能实现这些目标的行为，考虑产生可能实现它们的各种行为等。有些目标具有某种稳定性，人们可以带着成功的憧憬长时期地去追求它们。

XII. 个人会具有一些稳定的目标。

理性的人不仅考虑特定的(外在)结果，还会考虑他自己是个什么样的人；他会对自己成为不同样子的人之间有所偏好。Wp 表示当情形是 p 时，个人认为自己会成为的那个样子；Wq 表示情形是 q 时，他相信自己会成为的那个样子。(这些包括 p 或 q 将引起、塑造和促使行为人将成为的各个样子。)这里有一个可以为理由所压倒的前设，即对作为何种人的偏好要优先于各种较低层次的个人性偏好。(个人性偏好是指仅从对自己利益的估算中所得出的偏好。)

147

① 参见 Bratman, *Intentions, Plans, and Practical Reasons*。

XIII. 个人如果偏好 Wp 于 Wq,那么(所有情况相同)他不会持有 q 胜于 p 的(个人性)偏好。[①]

条件 XIII 认为个人是什么样子,他是何种人,在他的偏好中比他的个人性偏好具有更大的分量(要不然还能是什么情况)。(这一条件要受约于文化,只对处于某种文化里的人来说才是讲得通的吗?)

荷兰赌论证是某人的概率式(probability)信念应该满足概率论公理。它认为,如果个人的概率式信念不满足,且她总是愿意依据这种概率式信念来赌博的话,那么其他人就可以精心安排让她必输无疑,由此得到一个不可取的选项。这个论证认为,如果她的(概率式)信念是不合理的,那么她肯定会得到一个更低的效用数额。我们可以弄出这个论证的双重性,作为一个条件而施加:

XIV. 个人不应有这样的欲望,即若按它们行事,其结果就是得到不合理的信念或概率。[②]

这个条件会禁止有各种各样的东西:诸如,无论证据如何都想要相信某事;想要毫无戒备地与一个众所周知的撒谎者共处;想要自己(用酒精、毒品或无论什么东西)处于这样一种状态中,那里会**持续不断地**损害自己信念的合理性。但这个要求这样来陈述的话就太强硬了;或许相对于避免(某种特定的)不合理的信念或概率而言,按照这个欲望行事会为她带来更大的(合

① 把这个条件倒过来就太强了;并不是我们持有的每个偏好都蕴含这一点,即只要它被满足,我们就变成了另外一个人。

② 请注意,这一条件并不排除那些肯定会导致错误的甚或不一致信念的欲望。(在第三章的"合理性规则"一节中,我们说过,一个导致一组不一致信念的程序未必就是不合理的。)如果**对于**这种信念具有一种欲望的话,那就是另外一个问题了。

法)价值。① 类似地,荷兰赌要求通常陈述得太强了,也可能这世上会有这样的情境,那里持有那些不融贯的概率能带来更大的收益——比如某人会因为那些不融贯的概率而给予你一大笔奖赏。荷兰赌论证指出,肯定遭受损失,尽管如此,这一损失是可以被抵消掉的;因此,你违背条件 XIV 而得到的不合理信念或概率也是可以抵消掉的。为了避免这点,荷兰赌的寓意就不能说得过强,对于条件 XIV 也是如此。

这十四项条件在超越休谟而朝着对偏好与愿望的实质性约束这点上,已经带领我们走了一段相当可观的路程了。关于满足合理性条件和做出偏好选择的实际先决条件的——由条件 III、V、X 所要求的——经验信息,可能需要个人的偏好与愿望的十分具体的实质性内容;当与其他条件的约束相组合时,情况就更是如此。 148

我们能向具体内容更进一步吗?有一条激发性的思路,它试着把我们对信念合理性的看法与对欲望的看法相提并论。例如,人们一直认为,由于可靠的过程操作会产生高比例的真信念,所以若信念是由可靠的过程产生的,那么它就是合理的。细节的确会更为复杂,但我们仍希望能把这两者的复杂性相提并论。因而,按这种思路来看,合理的欲望便是由可靠地产生高比例__欲望的程序所形成的欲望。但我们准备如何来填这个空呢?对欲望来说,什么对应于信念情形中的真理呢?至少我目前还提不出任何独立的实质性准则。

然而,我们可以运用我们既有的条件,运用其他任何类似的条件,来规定该程序的目标:仅当偏好或欲望是由可靠地产生满足偏好结构的那些先决条件,即条件 I-XIV 的偏好与欲望的

---

① 感谢 Gilbert Harman 为我指出了这一点。

过程所形成的时候，它才是合理的。这不仅仅是说十四项条件应当得到满足，因为，任何能可靠地满足这些条件（我们能遵循）的程序，可能还会更进一步地约束个人的愿望和偏好。

XV. 一个具体的偏好是合理的，仅当存在得到欲望与偏好的程序 P，且

(a) 该偏好或欲望是由程序 P 所产生的，且

(b) 该程序 P 能可靠地产生满足上述（I 至 XIV）规范性结构条件的各种愿望与偏好，且

(c) 不存在任何更严格的过程 P'，以致该欲望或偏好由 P' 产生，且 P' 往往产生不能满足 I－XIV 条件的欲望和偏好。①

如果我们说，若偏好与愿望能满足条件 I 至 XIV（及类似的条件），则它们就是合理地融贯的，那么条件 XV 说的是，只有当偏好或愿望是合理地融贯的，且它是**由**一个能生成这种合理地融贯的偏好或愿望的**程序**所产生的时候，它才是合理的。

程序 P 不仅可以可靠地产生合理地融贯的偏好和欲望，而且其**目的**就能在于这样的偏好和欲望，还有能力把偏好和欲望塑造成和引导为合理地融贯的。程序 P 可以成为一个自我平衡机制，其目标状态之一就是偏好与欲望是合理地融贯的。在那种情形下，偏好与欲望的一种**功能**就是成为合理地融贯的。（同样，如果信念形成机制 B 旨在生成能接近于真理的信念，那么，信念的一种功能便是要成为近似地真的。）

149

因此，我们可以再加上下述条件：

XVI. 生成偏好和愿望的过程 P 旨在使它们成为合理地融

① 条款 c 是用来处理参照组问题（reference class problem）的，并排除那些经由一个会可靠地违背上述规范条件的亚程序（subprocess）所形成的愿望。没必要重视这里所建构的条款 c 的具体细节；它不过是承认有这样一个问题，是处理此问题的恰当条件中的一个定位器（placeholder）而已。

贯的偏好和欲望;P 是一个自我平衡机制,其目标状态之一就是偏好与欲望要是合理地融贯的。

类似地

XVII. 生成各种信念的认知机制 B,旨在使这些信念满足特定的认知目标,例如(接近)真、富有解释力等。B 是一个自我平衡机制,其目标状态之一就是要使信念满足认知目标。

偏好与愿望的一个**功能**是要合理地融贯;信念的一个**功能**是要满足认知目标。如果这些机制 P 与 B 果真是这样一种自我平衡机制的话,那么这就会从我们早前的功能论说中得出来。假定这些自我平衡机制确实生成了具有上述功能的信念与欲望,那么这样做是**它们**的功能吗?这取决于何种其他机制生成且维续了这些欲望与信念的形成机制。假如偏好与认知机制 P 与 B 本身是由另外的自我平衡机制所设计、生成、修改和维续的,且这些机制的目标之一就在于使 P 和 B 成为能生成合理地融贯的偏好(及近似于真之信念)的机制的话,那么这里就有了一种双重性功能。其一,偏好和信念的一种功能是分别成为合理地融贯的和近似于真的;其二,生成这种信念和偏好的机制的功能就是产生具有那样功能的那种东西。

XVIII. 存在一种自我平衡机制 $M1$,其目标状态在于偏好机制 P 产生合理地融贯的偏好,且 P 是由 $M1$(通过对这一目标状态的追求)所生成或者保持的。

XIX. 有一种自我平衡机制 $M2$,其目标状态在于信念机制 B 生成满足认知目标的信念,并且 B 是由 $M2$(通过对这一目标状态的追求)所生成或者保持的。

一种可行的想法是,我们的欲望形成机制与信念形成机制经历了进化和社会的塑造,其目的在很大程度上就是让这些机制具有这些功能。不止如此,一旦人们自觉到了他们的偏好与

150

信念,则他们就能指导这些偏好和信念,并监控它们对合理融贯性和真理的偏离,并做出适当的纠正。自觉意识(conscious awareness)成为了 P 程序和 B 程序中的一个**部分**,并且它还有意识地把这些程序的目标确立为合理融贯性和真理。

XX. 自我平衡的偏好和欲望形成机制 P 的一个成分就是人有意识地力图得到合理地融贯的偏好和欲望。

XXI. 自我平衡的信念形成机制 B 的一个成分就是人有意识地力图得到那些满足认知目标的信念。

这种自知性和监控使我们得到了一种更丰满的合理性。(有人可能表示只有当这些条件被满足时,我们才具有合理性。)

自觉意识不仅能监控那些偏好和信念,而且还能监控据之产生它们的那种程序,即 P 和 B 本身。自觉意识能够修正并改善这些程序;它也能够重塑这些程序。因此,自觉意识就成为 $M1$ 和 $M2$ 机制的一部分,由此能够起作用来确定那些偏好与信念形成机制本身的功能。

XXII. 维持 P 的自我平衡机制 $M1$ 的一个部分就是人们有意识地力图使 P 产生合理地融贯的偏好。

XXIII. 维持 B 的自我平衡机制 $M2$ 的一个部分就是人们有意识地力图使 B 产生满足认知目标的信念。

由此,合理性就得以塑造与控制其自身的功能。[那么,自知(self-aware)合理性是否也会考量那些产生和维持 $M1$ 与 $M2$ 的程序,并因而也在那些程序中发挥某种作用呢?]

另外一种可能性是一种历史理论。合理的欲望是通过某类(很可能合理的)程序从(或能从)一系列原始(生物性的)既定欲望中派生出来的。那么,欲望的合理性就是相对于生物性起点的(和派生或转化过程的);但外星生物(Alpha Centaurians)可能就会起始于与此迥异的先天欲望和强化刺激(reinforcers)。

但是，我们生物性给定的欲望是合理的，这点应该据定义就是如此吗？这一点或许标出了我们合理性的一个局限，亦即我们都是**造物**。我们开始就有某些欲望和倾向；尽管我们不会永远分毫不差地固着于此——我们能够以各种方式修正和转变它们——但我们总是从**那**而来的。 151

我要再次强调，我讨论这二十三项条件的目的不是要认同这些**特殊**的条件，也不是要捍卫其特定的细节，而是力图展现出这类合理性条件具有何种空间，并表明当这些条件摆到一起时，休谟式画面是如何得到改变的。

可检验性、解释和条件化

还有两种路径可把更多的内容归于合理的欲望和偏好，但我对它们疑虑从生。第一种路径追问，若决策理论要成为一种具有可检验内容的经验性理论，亦即人们发现它有被违背的可能性，那么有什么是必须成立的呢？① 如果结果得到细致的解释、拆分和描述的话（诸如，星期二下午 5 点收到 x，当从某个生于这样一个日期的人手上所拿到的），那么实际的选项的任何模式都可以构想成符合这些规范条件。若把特定细节看作固定的会导致违反规范性条件，那么就对选项进行更细致的描述（这总是可以一直做下去的），从而消解这种偏差。论证继续进行：巩固可检验性的这种方法就在于引入选项各个方面的完全清单，这些方面是理性人会考虑的或会放任这些方面影响她的选择。

① 参见 John Broome, *Weighing Goods* (Oxford: Basil Blackwell, 1991), pp. 100 - 107; Susan Hurley, *Natural Reasons* (Oxford: Oxford Univ. Press, 1989), chs. 4 - 6。

然后,我们就可以检验:她做出的选择是否满足针对备选项的这些规范性条件,这些条件是由可合理地加以考虑的**那些**方面所确定的。因此,若决策理论要成为一种经验的且可检验的理论,那么它也预设了合理的偏好,或者至少是有进一步(合理)内容的偏好。

我认为,这种论证推进得太快了。对选项的**任何**规定(亦即用来区分各选项的那些方面)都会把决策理论规定为一种可检验的经验理论。该规定不一定列出那些被认为可合理地考虑的各个方面。尽管如此,我们还是能够发现个人一直在违背针对**那些**特定选项的条件。的确,此人的捍卫者也许会主张该人**是**满足对于其他选项集的规范条件的。现在我们来考虑这种存在量词(existentially quantified)假说,亦即存在**某个**选项集(是由某些方面的集合所确定的),个人对它们是满足规范性条件的。如果只作为对她既往选择的一种描述,那它是成立的;总有这种或那种规定会契合于他的既往选择。但这并不意味着,该规定就一直在**指导**他的选择。如果偏好(至少)是一种选择倾向,那么宣称对那些规定的选项有那些偏好的话,这就使他承诺了进一步的后果,不仅仅有关他实际的未来选择,而且还有他**可能**做出的那个特定的其他选择——我们一直假定他的偏好保持不变。这一假说有着诸多假设性结果。这不是一个无足轻重的真理或说逻辑真理,而是一个经验主张,存在规定各种选项的方面集,且个人确实按照和将按照对于这些选项的规范性条件做出选择。也有可能根本就不存在任何这样的方面集。①

152

---

① 请注意如下二者间的并列性:其一是对决策理论的可检验性的讨论,我们将其解释为一个存在量词(existential quantifier)("这里存在着一组方面来规定各个选项,以使……");其二是我们先前对"适应性"问题的论说,它对可遗传的基因组特征采用了一种存在量化(existential quantification)处理。

但是我们要如何来发现它们是否存在呢？规定出这种具体方面的负担，亦即陈述那些存在量词假说的存在范例的负担，是落在倡导这个主张（即特定人的选择要满足这些条件）的人身上，还是落在否认它的人身上呢？相对于较容易地规定一个存在量词陈述的范例而言，确立否定掉这样一个存在量词陈述的负担大得多，人们可能会认为规定的负担要落在声称规范条件要得到满足的人身上，尤其当规定似乎不同于通常的及显而易见的那些规定时。但是，就我们正在考察的这种论证而言，我们只需要注意，对于备选项的任何一种规定都要产生在原则上可检验的假说。这不一定要以特别倾向理性的方式来规定这些选项。考虑这样的情境，个人对任何选项*不*满足契合于所提出的合理方面清单的条件，然而他满足*其他的*（非合理性）方面清单条件。（也就是说，在过去的选择中，该人是满足*那些*选项的条件的，亦即规定已经提出来了；如今我们是检查这个人未来的选项，并去发现他确实继续满足那些规定选项的条件，现在的规定方式不是一种对已观察数据的特设描述，而是事先提出的规定，要由其预测的准确性加以证实。）我们所考虑的这种论证的倡导者真正想说的是：若这个人没有按照决策理论行动，这就是因为该行为人对选项的规定中并不具有他们（或许是正确地）认为是合理内容的东西吗？

涉及更多合理内容的第二种论证是一种诠释性(interpretative)论证。除非我们赋予个人的欲望和目标以某种内容，以某种使得它们类似于我们自己偏好的（即类似于我们在他们的处境下会有的偏好的）方式来界定它们，否则我们不可能把人诠释为在做意向性行动，即具有偏好和欲望。无论如何，这是晚近流行的各种不同的厚道(charitable)诠释原则的主旨，首先是用在陈述和信念的内容上，后又被更广泛地应用于它们与

153

偏好的组合上。[①] 我发现把诠释的这种指导原则即便应用到陈述和信念都是不恰当的。我会在它们最强的这个语境中，而不是在把它们进一步应用于偏好和欲望的语境中展开对它们的讨论。(它最近在哲学领域内的其他论题上的应用具有极大的重要性，即使只是因为这点，这种观点就值得我们好好讨论。)

我不满的根源在于，这些立场看来是帝国主义式的，它们对我们碰巧具有的立场，对我们自己的欲望和偏好赋予了不恰当的分量。人们也许试图假定我们都具有一个共同的进化背景，从而避免这一点。不过，这样做会限制这个主张的一般性。迄今为止，这个主张一直是针对地球或宇宙其他地方的所有理性行为者的主张。由于没有任何单个欲望是服务所有环境之下的全面适合度的必要条件，所以看来并不存在一种富有阐明性的一般性层次，使得我们可以预测*每*个进化生物都共有*某些*欲望。若我们的信念或偏好对我们自己都缺乏权威，则它们通过这种诠释论证而赢得权威，使我们自己的信念与欲望成为一切信念与欲望的基准，这看起来就是讲不通的。我们对怀疑论的一般根据就说这么多，现在转入细节。

把个人的所说所写尽可能诠释为或翻译为*真*，这样做是没用的，因为我们知道此人无从得知我们所谓的真理是什么；他未曾做过试验、收集数据等。更有道理的也许是

1. 诠释或翻译个人所言所行的方式要尽可能使个人显得

---

① 参见 W. V. Quine, *Word and Object* (Cambridge, Mass.: M. I. T. Press, 1960), pp. 57 - 61; Donald Davidson, *Inquiries into Truth and Interpretation* (Oxford: Oxford Univ. Press, 1984), 论文 9 - 13; David Lewis, "Radical Interpretation", 载于他的 *Philosophical Papers*, Vol. 1 (Oxford: Oxford Univ. Press, 1983), pp. 108 - 121; Ronald Dworkin, *Law's Empire* (Cambirdge, Mass.: Harvard Univ. Press, 1986), ch. 2; Hurley, *Natural Reasons*, ch. 5。

是理性的。[1]

这个建议含有这样一项假定,此人与我们都使用着同样的合理性标准。然而有很强的初定证据表明,有些人并不是按照我们所接受的那种正确标准来推理或推论的。[2] 亦有证据表明,某些毒品能使人的信念形成过程变得不那么合理。有些人明确宣告,他们遵循的是不同的合理性原则。[3] 因此,若把我们的合理性标准归于他们,看来就会扭曲这些情形。不仅如此,社会的合理性标准也会随时间而变化和发展。一直有人主张:识字促使人们关注一致性,关注有理据的支持——因为人们可以对书面陈述反复检查且与其他陈述做比较——由此导致构建出新的推理与批评标准。[4] 因此人类学家要警惕,不要把某些合

154

① 对这种提议的各种版本,请参见:David Lewis, "Radical Interpretation", pp. 108 – 118; Richard Grandy, "Reference, Meaning and Belief", *Journal of Philosophy* 70 (1973): 439 – 452。

② 参见 *Judgment under Uncertainty: Heuristics and Biases*, ed. Daniel Kahneman, Paul Slovic, and Amos Tversky (Cambridge: Cambridge Univ. Press, 1982); Lee Ross and Richard Nisbett, *Human Inference* (Englewood Cliffs, N. J.: Prentice-Hall, 1980); and Paul Thagard and Richard Nisbett, "*Rationality and Charity*," *Philosophy of Science* 50 (1983): 250 – 267, 它讨论了对于表述一种解释原则的那些心理学结果的含义。但也要参见 Tversky and Kahneman 在 Gerd Gigerenzer 所做调查中的相对照的观点,"How to Make Cognitive Illusions Disappear", European Review of Social Psychology 2 (1991): 83 – 115, 并且也见于 Leda Cosmides and John Tooby, "Are Humans Good Intuitive Statistics After All?" (即出)

③ Paul Thagard and Richard Nisbett, "*Rationality and Charity*,"其中举了禅师(Zen masters)和黑格尔的例子。

④ 参见 Jack Goody and Ian Watt, "The Consequences of Literacy", *Comparative Studies in History and Society* 5 (1963): 304 – 345; Jack Goody, *The Domestication of the Savage Mind* (Cambridge: Cambridge Univ. Press, 1977), pp. 36 – 51, 74 – 111; Goody, *The Logic of Writing and the Organization of Society* (Cambridge: Cambridge Univ. Press, 1986), pp. 1 – 20, 171 – 185。

理信念的标准施用于那些没出现文字的社会,因此也不要以那里的个人满足这些标准的方式来诠释他们。

退一步的立场可以认为:成年人例示和展示出不同的合理性标准,这很可能是因为不同的文化和他们自己的个人生活经验所导致的,而所有的婴儿开始都具有相同的信念形成模式,亦即形成信念的相同能力和禀赋。形成信念与欲望的那些起始过程会处理此人的生活和文化给予她的那种经验和数据,所有成熟的信念都是这种处理的合理结果。即便在信念形成和获得新倾向的倾向方面,个人之间不存在任何初始差异,他们生活经验上的差异仍然不仅会带来具体信念和欲望上的差别,而且会在他们的信念与欲望获得模式和程序上,还有如何按照它们行动上,造成差别。有些不同的程序不仅仅是按照先验程序从个人经历得出来的,还有可能是经由不同的经验所**植入**的。科学的理论应能解释由此导致的所有结果上的差异(不管存在什么样的初始差异),尽管这并不是说,某种合理性理论将能应用于每一个阶段或任何阶段。(**合理的**这一术语可能无法应用于普通的初始儿童阶段,这阶段是积累信念的。)

戴维森(Donald Davidson)主张,我们无法融贯地理解"极端不同的概念体系"这个概念。他的"极端不同性"的准则是令人吃惊的不可译性(nontranslatability)准则——这与我们通常使用的"不同的概念框架"这个概念不一致——但尽管如此,他的主张看来仍然蕴含了,在每种语言文化中,都必定有某种共同的合理性(strand)在起作用。① 但戴维森本人却为他自己的观

---

① 参见 Donald Davidson, "On the Very Idea of a Conceptual Scheme", 载于他的 *Inquiries into Truth and Interpretation* (Oxford: Oxford Univ. Press, 1984), pp. 183 - 198。

点提出了一个反例：可译性在不同的概念体系之间可能不是传递性的。或许我们可以译土星语(Saturnians)，而土星人可以译冥王星语(Plutonians)；但我们就是无法译冥王星语，其概念框架是我们理解不了的。戴维森试图扭转这个例子的企图是软弱的，他问道：我们如何得知土星人是在**翻译**冥王星语呢？答案在此。他们当时做的是同样的事情，亦即，运用的是他们把英语译成土星话时所使用的相同的程序(和元程序)；这一程序是将每个英语句子 Ei 映射为土星语中的相同句子 Si，而 Si 也正是**我们**(在翻译时)映射为英语中之 Ei 的句子。正如我们有很好的理由认为我们是在翻译一样，我们有同样的理由认为他们也是如此。

155

然而我们不能把土星人已经从冥王星语翻译而来的土星语句再转译成英语，由此来翻译冥王星语吗？在这种情况下，我们归根结底不是理解了冥王星语吗？可以主张，即使居间的土星人并不存在，但他们存在的可能性本身就足以表明，我们是有可能懂得并且合理地翻译冥王星语的，而这就是戴维森论点所需要的一切。① "直接可译的"观念是非传递性的。但当两个不是直接可译的群体或个人，处于经由居间物的直接翻译链条上，这难道不是足够构成间接翻译，由此是翻译**本身**了吗？我并不这样认为。

"直接可译"是非传递性的，理由在于它(仅仅)要求概念或者陈述的应用条件之间有足够大的重合，而不是要求它们完全重合，这样一来，一个有着足够重合的链条就可能导致链条两端之间没有足够的重合。因此我们可以设想，有两条串联着不同星球文明的不同的直接可译性的居间链条；链条的每个环节都

---

① 参见 Susan Hurley, "Intelligibility, Imperialism, and Conceptual Scheme", *Midwest Studies in Philosophy* (即出)。

是足够重合的。我们从源点 S 开始以很大的重合逐渐转移的一个链条，把我们最近的链条邻居 Y 的陈述 y 翻译成 z；通过从同一个源点 S 开始的另一条有很大程度重合的逐渐转移链条，我们把最紧邻的另一个邻居 X 的陈述 p 翻译成了 q。然而，z 和 q 之间的差距，可能你要它有多大，它就有多大，就像“母牛”和“熵”这两个概念那样是风马牛不相及的。① 因此，我们不能通过设想存在某个居间的翻译链条来论证相互之间可理解性。他必须能够表明，不可能存在两个这样不同的链条，它们导致了极端不同的翻译。然而，很清楚，这种不同的链条是可能存在的；而且一旦我们知道了它们，我们难道不是会认为我们根本就无法理解端点上的那种文明吗？②

156 我们也许还有办法把自己投射进完全不同的概念框架中去，例如，我们尽管不使用翻译，但有可能使用能制造奇幻体验的某些药品或程序，从而进入奇幻的体系。麦金泰尔(Alasdair MacIntyre)认为，我们可以像另一个社会的一个小孩子那样，从头开始学习当地的习俗和语言，然后我们发现无法将那里的言

---

① 这里有一个类比。在平行分布处理系统中，当次区域有一系列的交叠区域时，有一些空间代表了中间单元被激活的向量，其中的每一区域都与另一区域足够地接近，可以作为对该区域中任何特定向量的一个吸引因子(attractor)，但是第一个和最后一个区域可能是极为独特和不同的，亦即接近度还不到能吸引链条上另一端区域中的向量。这样两个链条会趋向不同方向，导向非常不同的端点区域。对于生发出这两个不同链条的那同一个区域，或许无法用我们自己的语汇进行概念化描述[只有能确切表明权重与活动向量的一种语汇才能做到]，即使这两个链条的终端区域确实具有这种概念化描述，它们也是彼此完全不同的。

② 或许我们能找到某种理由认为，这些长链中的某一个要比另一个链条更加蜿蜒、更加曲折，因此假设情况不是这样：在很大的概念空间中，每一条路径的蜿蜒度都是等同的。下面可以通过思考这一问题而继续我们的讨论：我们如何知道链条上的每一个环节(即使是那些远离我们的环节)都包含了一种翻译呢？那个概念必须是直接地足够近似于我们的概念呢？还是间接地翻译它就够了呢？

语翻译成英语。[1] 因此,我们也许能够使用其他的概念框架,尽管根据我们的标准或我们能够表述的任何其他标准,它都不是合理的。

德沃金(Ronald Dworkin)提出了诠释的另一种指导原则:

2. 这样来诠释(和翻译),使得此人的言行能尽可能地好,即尽可能地有道理、有原则和正确。[2]

德沃金的这个原则,首先是在制度的语境下将此作为诠释规则而提出的,但后来他认为对于文学解释也是可行的。在这种诠释模式下,事物是“从最好的角度”(亦即有可能的最好角度)来看待的。我们可以把这种对每件事物涂上最光鲜表面的原则称作极度乐观原则(Panglossian Principle)。德沃金建构出的这个原则,使得这一态度不必会看作导向种族灭绝或者奴隶建制。但(就我们所知)有的制度代表了在相互竞争的各种利益之间的*妥协*,有的制度是由试图攫取与维持权力而来的混乱结果所塑造成的,在这样的情形下,这个原则难道不也是误导性的吗?如果我们不是透过一个意在合理化的多棱镜来看待这种制度,而是将它看作实际产生它的那一程序之结果的话,我们不是能更好地理解它吗?

让我们考察另一项指导原则:

3. 这样来诠释(或翻译),使得此人所说的尽可能地好理解(intelligible)

这一原则也为其他的指导原则开辟了空间,因为真理、合理性和善性(goodness)都是可理解的*方式*。但它们并非是惟一的

---

① 参见 Alasdair MacIntyre, *Whose Justice? Which Rationality?* (Notre Dame, Ind.: Univ. of Notre Dame Press, 1988)。

② 参见 Dworkin, *Law's Empire*, pp. 46-68, 76-86。

方式，偏离这些方式也是可理解的。然而，我们并不总是想使得所说之内容尽可能地好理解。这个人可能吃了药，而我们知道这会让他胡言乱语。

原则 3 因此必须被修正为：

157 4. 这样来翻译“p”，使得在该语境下该人说了“p”这个事实尽可能地好理解。

要使得其可理解的是整个事实，亦即该人在该情形下说了 p。原则 4 就如何翻译 p 为我们提供了指导，而且除了可包含人们的目的与动机、惧怕与迷信外，它还有余地来包括那些历史学家和人类学家在考虑历史以及文化语境时所做的事情。

我们还可以采取更宽泛的视野。行为人不仅说了 p 而且还说了 q、r 和 s。那么，我们难道不应该要这样翻译，使得该人在不同语境之下所说的所有那些具体事情都尽可能地好理解吗？他的同辈人以及前辈人也说过其他的东西。那么，难道我们不应该要这样翻译，使得他们所言的诸事实都尽可能地可理解吗？而这不是也许会颠覆或者修改那种使得他说 p 这一事实（独立处理下）尽可能地可理解的东西吗？

先前所述的那些原则——最大化真理、合理性或善性质——是激动人心的，但还是不够的。不过原则 4 看来是显而易见的，但也很乏味，没有阐释力。“当然，我们确实是这样翻译的，使该言论（亦即所说的）尽可能地可理解。”

我们让他的话对谁尽可能地好理解呢？这是对我们自己，也就是翻译者和解释者而言的。因此，当我们判断何种解释使得事实最可理解时，我们自己的可理解性标准就进来了——与我们关于言论、行动、生活和社会的理论一起。然而，我们的标准却并不能灌输进他的观点，也不能归于他——尽管使他的观点尽可能地合理或好的这些原则正是在这样做。“言语含混”这

一事实是可被理解的，但不必要求其内容也是可理解的。使得这个言说可理解，而所说的内容并不(必然)是可理解的。

个人说他曾做了什么，要使这个说的事实尽可能好理解，就是要尽可能好地解释他对此的言说。这就会用到我们关于人类行为的理论，用到我们有关它的常识，且把他的言说契合进来。我们的解释网络越详密，他的言论契合于这个网络的细节越多，那么我们就使得这些东西越可理解。要表明“他言说所做之事”的目的，也就是说要确认，他在言说时所做的是何种以言行事(illocutionary)的行动——这一切都能增加他言说那事的可理解性。

换言之，遵循前面讨论的指导原则，使得他说的东西尽可能地真、合理或善，我们是否有助于使得他言说那事尽可能地可理解呢？只有当我们有理由认为当时他是准确的、合理的或善的时候，情况才是如此。① 如果我们有理由认为他*不是如此*，那 158

---

① 几年前，William Dray 在反对 C. G. Hempel 将解释的(演绎或统计)规律模型(nomological model)作为历史解释的一个必要条件时，就认为有这样一种历史解释模式，即合理解释，它不利用法则来解释一个行为，而是将该行为解释成在那些环境下要做的(一件)合理的事情。参见 William Dray, *Laws and Explanation in History* [London: Oxford Univ. Press, 1957]，Hempel 对此回应道，在这种合理解释模式中，我们需要再添加一个条件，即行为人具有合理地行事的倾向，并在那时会运用这一倾向，亦即这是一个他会合理地行为的情境。因此，Dray 的解释模式中包含着一条隐性法则：对于该行为人和一定范围的 *s* 类情景或环境而言，他会在一个 *s* 类情景中合理地行为。这一解释模式还可以继续：该行为人处于情景 *si* 中；*si* 属于 *s* 类情景；在情景 *si* 中，行为 *A* 是要做的合理的事情；由此，他做 *A*。(参见 C. G. Hempel, “Rational Action”, *Proceedings and Addresses of the American Philosophical Association* 35[1961 - 1962]: 5 - 23。)对某个个体而言，最需要解释的恰恰是他为什么当时会合理地行为？既然他做事几乎总是不合理的，因此仅仅说他当时所为乃是要合理地做的事情，这是不够的。这是碰巧的吗？他是否有在很少的特定情境中合理地行为的倾向，而这个情境恰好是其中之一？他是否当时有一种突发的合理性呢？

么，将他的话解读成准确的、合理的或善的，这就只会制造出迷惑，而不是可理解性。

然而，假定某个人打算使自己所讲的内容——不是他言说这件事情这一事实——尽可能地可理解。那么这能否是诠释的另一个目标，它是否包含使该陈述是真的、正确的、有道理的或合理的呢？然而，即使是谎言、邪恶命令和侮辱的言辞都是完全可理解的——也就是说，它们的**内容**乃是完全可理解的。我们可能恰恰不能理解的是，此人何以会讲或颁布这些东西，而这点又把我们带回到诠释这些内容的言说之上了。

尽管如此，当我们看到对某事所能给出的最好理据（无论说它的那个人是否能够或会提出这个理据），我们看到**能够**说什么来支持它时，我们不是能够更好地理解它吗？我并不否认这一点，正如当我们看到有什么能够反对它时，我们也能更好地理解它一样。在发现它的优劣，即揭示出它与其他理论或问题的关联时，我们能够更彻底地探讨其本质。因此，最厚道的诠释方式只是展现出了事物本质的一个面相而已。

请注意这样一些类比，我们一直讨论的问题和有关进化理论最近的讨论之间的类比性，还有两类人之间的类比：一类人是为大范围的特征给出适应解释（即把它们看作一直是进化选择的），还有一类人把许多特征不解释为具有直接的适应价值，而是将其看作由其他进化方式所生成的，例如，作为**被**选择的其他特征的副作用，或者是作为满足各种结构性约束或基因漂变的结果等。认为理解一部知识产品（intellectual product）的**唯一**方式就是使其尽可能地合理或好，这类似于认为：解释某项生物特征的唯一方法就是表明它是最优的，或者有最大适合度的。我们使该知识产品的产生能尽可能可理解的这一原则，对应于这样一种观点：我们承认有诸多不同的进化解释模式，而

159

把何种模式适用于既定的某一(类)情形看作是个经验问题。[①]尽管这两个问题在结构上是并列的,人们还是可能在这两个不同的领域中追求不同的解释策略,也许是因为人们认为在生物领域中的选择压力要更大一些。让我提到另一种可能路径来深化偏好内容的条件。对概率进行历时性修正的贝叶斯条件已经提出来了,即所谓的贝叶斯条件化(Bayesian conditionalizing),人们也许会把这些与效用的跨时条件(intertemporal conditions)相提并论。假说 h 基于事件 e 在昨天的条件概率被假定后,那么当今天已知证据 e 存在时,我们就可以基于先前的那个条件概率去求 h 在今天的新概率,同样,人们由此也许会发展出一个条件效用的观念,并提出一种类似跨时条件化的条件。[②]

然而,如果昨天的特定概率或偏好没有任何特殊的权威性——其要求的一切就是满足概率论公理或偏好的冯·诺伊曼-摩根斯坦的规范性结构条件——那么它们为什么应该对今天的特定概率与偏好具有任何权威性呢?偏好的结构条件在任何时刻都要求这些偏好必须以某类方式组织在一起;如果不是如此,那么何种偏好要受到修正就是开放的。偏好要具有传递性这一规范性条件,行为人偏好 x 于 y 且偏好 y 于 z 的前提,人们结合两者并不能得出:她应当偏好 x 于 z。或许她不应当偏

---

① 参见 Stephen Jay Gould and Richard Lewontin, "The Spandrals of San Marcos and the Panglossian Paradigm: A Critique of the Adaptationist Programme", *Proceedings of the Royal Society of London*, B 205 (1979): 581 - 598。Daniel Dennett 不但在诠释(interpretative)任务与进化解释任务间做了类比,并且坚称两者实际上就是同一个任务,因而必须都接受一种最优假设的指导!参见 Dennett, *The Intentional Stance* (Cambridge, Mass.: M. I. T. Press, 1987), pp. 237 - 321。

② 关于这种条件性效用观的早期例子,参见我的 *Normative Theory of Individual Choice*, pp. 144 - 154。

好 x 于 y 或者偏好 y 于 z。偏好应该是传递的这一要求不应当被解读为：如果个人偏好 x 于 y，并偏好 y 于 z，那么他应当偏好 x 于 z。相反，这个传递性要求要解读为：情形不应该是这个样子的，即一个人偏好 x 于 y，且偏好 y 于 z，但并不偏好 x 于 z。因此，从这一条件中，我们推导不出个人应持有何种具体偏好的任何超然结论。[①]

请思考偏好应该是传递的这个结构条件的一个(公认的)跨时版本：如果个人在周日偏好 x 于 y 并且偏好 y 于 z，而且他在 160 x 与 z 之间没有任何偏好，那么他在星期二应当偏好 x 于 z。这是一个可反驳的条件，因为或许在星期二这个人反而不再偏好 x 于 y 或偏好 y 于 z 了。当然，这个人在星期二并不能改变他在星期一时偏好 x 于 y 且偏好 y 于 z 的这个事实。人们能如做如下主张吗？虽然一度被正确解释的条件只是说，各种偏好不应以某种方式互相冲突，而哪些偏好要被修正以得到和谐则是未定的，尽管如此，一个跨时条件加上过去不可变更这一事实，现在还是可以得出结论：行为人今天应当拥有某一特定的偏好。这个主张同样是可反驳的。假设该行为人在星期二变成偏好 z 于 x 了，并且希望他自己在星期一时不曾偏好 x 于 y 或不曾偏好 y 于 z。因为他无法更改那些既往事实，那么这就意味着他现在必须建立起一个他同样会反对的当下事实，即偏好 x 于 z 吗？肯定不是如此。

然而存在着众所周知的吸钱机论证：如果个人拥有非传递的偏好，且他愿意付出少量的钱以实现他更为偏好的那个选项，那么他就会由一系列的支付而走完一个循环，最后又回到起点，

① 对于这一段的观点，参见我的 *Normative Theory of Individual Choice*, pp. 94－98。

只是更穷了而已。(而且这个循环还可以重复下去。)跨时情况也是如此,那个人在星期一偏好 x 于 y 且偏好 y 于 z,但到了星期二却偏好 z 于 x,如果星期一他在 z 的位置上,那么他可以导致付钱回老位置。(他会在星期一付一笔钱从 z 转换为 y,然后再付一笔钱从 y 转换为 x,接着在星期二他又会付一笔钱从 x 转回到 z。)然而这个事实并不能表明此人在星期二应当偏好 x 于 z,或者好像有这种偏好那样去行为;因为他本来就不是那样的,他也没有必要那样。他的偏好能够变化,并非只能是他的既往偏好含义的发展。跨时性规范条件规定人们不得推翻其偏好,这无疑是荒谬的,即便对于一个果真会连续每天付钱直至最后回到原点的人来说,情况也是如此。即使这是不对称偏好的跨时类比物,这个规范条件也还是完全讲不通的。

类似的反驳同样适用于条件效用概念,这种概念是为了进一步的跨时条件化所要求的——毕竟,实际上还无人如此提议过——而且,这些反驳也适用于广受赞同的下述标准:即概率应当依据贝叶斯条件化的原则而进行历时性变化。① 依据严格的贝叶斯主义者[对他们来说,惟一允许的概率就是测度人们信念度(degree of belief)的概率]的学说,这些个人性的概率只需要能满足概率论公理。任何具体的概率安排,即信念度的任何组合,只要满足这些公理就都是同样好的。② 因而,星期一的特定概率不具有任何权威性。对于任何时刻的概率为什么应该满足概率论公理,贝叶斯主义者确实有一个论证。偏好上的吸钱

161

---

① Gilbert Harman 曾独立地提出过反对意见,尽管他并没有从事跨时荷兰赌论证(intertemporal dutch book argument)的研究,但认为既往概率不应当具有约束力。参见他的“Realism, Antirealism and Reasons for Belief”(即出)。

② 严格的贝叶斯主义者还可能希望行为人在其特定的个人概率中显示出“良好的判断力”,但并没有更进一步的具体规范性条件来限定这种要求。

机论证相对应的是概率上的荷兰赌论证：如果个人持有的概率(亦即信念度)不满足概率论公理，总是愿意按信念度的机会来赌博，那么可以给出一系列他会接受的赌博，以至于所有赌博的结果是他赢不到任何钱，而且随我们所考察的公理不同，他要么是可能输掉一些钱，要么是必定输掉一些钱。①

这里同样还有一种跨时版的荷兰赌论证：一个不依据条件化原则而历时性地变化概率的人，他会接受给出的这样一个赌博，经过这个赌博他不可能赢钱，且最终必定会输钱。② 但考虑这样的一个人，无论出于何种原因，他决定持有这样的概率，这个概率与她昨天的概率并不处于想要的那种条件关系中。这样的一个人或许希望自己的概率能历时性地组合在一起，因此她希望自己昨天的概率不是那样的，只是为时已晚，无法修改了。那么现在的她是否应当被过去所拖累呢？亦即被那些除了当时曾共同满足了(非时间性的)概率论公理外无任何好处的概率所拖累呢？但那就要求她今天必须持有一个她不愿意持有的信念度，并且意味着她今天必须在一个她不愿意下注的机会上下注。(她不是应当把昨天的概率当作一种沉没成本吗？)无疑，她可以

---

① 当一般不是对每一命题 q 上都打赌时，在命题 p 上打赌这一事实 $F1$ 可能就会改变 p 的个人性概率，从而不同于以往。给定 $F1$，打赌的行为便标出了此人对 p 的个人性概率。假定不是另一个人提供的这个赌局，但是存在着一个可以做 A 的情境，其报偿实际上对应于在 p 为真上打赌的报偿。那么做这件事标明的便是，在给定 $F2$ 的情形下($F2$ 是指我们可以对 p“下赌”，但不是对每个命题 q 都可以)，行为人对 p 的个人性概率。我们在何种条件下才能得到以前的那种概率，它既不受传染，也没被改变呢？

② 参见 Paul Teller, “Conditionalization and Observation”, *Synthese* 26 (1973): 218 - 258, 陈述了一个归因于 David Lewis 的论证。Bas Van Frassen 注意到这个论证取决于违反条件者遵循了其他某个特定的规则；因此他主张，违反条件化是可以的，只要没有遵循某种规则。Bas Van Frassen, *Laws and Symmetry* (Oxford: Oxford Univ. Press, 1989), pp. 160 - 176。

依据今天的概率行事,并为自己昨天没与别人打赌而大呼幸运。但即便她昨天果真打了一些赌,那也并不意味着她今天就必须为她现在所反对的、昨日的那些信念而打赌,继续把钱扔到水里。假设她把昨天的概率视之为在发烧状态下的产物——现在意识到了那是幻觉——或她相信当时是上帝直接对她说的信念。那么贝叶斯主义者还坚称她今天必须以那些概率为条件吗?或者,贝叶斯主义者是否会认为,昨天的概率并非是以前天的概率为条件的,因此那些概率在今天可以被忽略不计吗?在星期二的时候,行为人并没有以星期一的概率为条件;那么到星期三时,她是应当以这些错误地得出的星期二的概率为条件,还是应根据从星期一起所获得的全部新证据以星期一的概率为条件呢?她还可以因为想出了与既定假说不同的新假说,从而拒绝自己的旧有概率;为"新假说"范畴中加入更多的结构,明白此假说的具体细节可以导致她对定义域重新概念化,据此以一种更近似于重新分派初始概率的方式来分配概率,而不是在新信息的基础上来条件化既有概率。① 162

即使她昨天确实依据当时的概率打了赌,她现在依然可能想要按照她今天的概率判断来下注。的确,她这样下注将必输无疑。但是,或许相对于一个可能的重大损失来说,她更加偏好一个确定会发生的小损失。而根据她今天的观点来看,基于以昨天的概率为条件来下注(这在她今天看来是更差的机会),前者便是将会发生的事。

然而,我们对偏好的立场并不同于严格的贝叶斯主义者对

---

① 参见 John Earman, *Bayes or Bust* (Cambridge, Mass.: M.I.T. Press, 1992), pp. 195-198。对于贝叶斯式条件化的进一步批评,参见 F. Bacchus, H. E. Kyburg Jr. and M. Thalos, "Against Conditionalization", *Synthese* 85 (1990): 475-506。

个人性概率的立场；他们的概率只要求满足概率计算的公理。而在前一节中，我为概率提出的条件已经超出了通常的结构性规范条件。若该人昨天的偏好与效用满足了这些进一步的条件，那不就足够堵住“或许他昨天本来应该持有不同的偏好”这样的话语，还可证成对他以当时实际持有的那些偏好与效用为条件了吗？然而还存在着其他的一些偏好，它们也是可以满足这些结构性规范条件和那些进一步的条件的。全部的要求似乎只在于，行为人今天达致某一套令人满意的偏好，亦即达致一套能满足所施加的全部条件的偏好，只要这些偏好原本可以用昨天的这套(或其他)令人满意的偏好为条件而形成。但后面这套偏好不一定要是他昨天实际持有的那套偏好。而且，这一项跨时条件很弱，似乎并未施加任何更多的内容于此前的要求上——亦即要求个人的偏好在某一时刻是令人满意的，这就是说，能即时地满足此前所提出的关于偏好的结构性条件和额外的条件。

在我看来，我们在本节中所考察的所有路径，都不能更进一步地对偏好产生出恰当且内容更丰富的条件。我确信，我们应当超越严格的休谟式观点，而且我在前一节中也暂时提出了一些实质性条件。我本人欢迎一个更强有力的、内容更丰富的有关偏好与欲望之合理性的理论。除了这里已经说过的之外，它还能进一步约束与限制它们，即便无法完全确定它们。

163

康德在遵从原则的伪装下，试图从合理性概念本身推演出目标来。我们在前面主张，原则是具有特殊功能的，亦即被设计成配合(且修正)既定目标而发挥作用的工具。拆开原则与任何既定目标之间的关系，然后仅仅从恪守原则之观念中推演出目标，或推演出对行为准则的特定约束，这样的康德式企图是会失败的。如果我们的合理程序是被设计成与既定的生物性目标

[或者是与我们的欲望和目标,而在我们的进化史上,对它们的追求乃是(在统计学意义上)与全面提高适合度相关联的]相配合而起作用的,那么,毫不奇怪,想合理地从无(*de novo*)(也就是不以**任何**目标或欲望而开始)来推出目标的企图一定会失败。这并不是说我们就固着于那些欲望和目标。我们的合理程序使我们能够在相当大的程度上修正它们。其步骤是做出单个微小的变化,反复地这样做,然后就会产生很大的累积性变化。

假设某人真的成功设计出了欲望的一种完全恰当的实质合理性理论。给定这样一种理论,那么我们就能说,若塑造信念的过程或行为在达致**合理的**(而非任何被专断地给出的)目标时是**工具地**有效的,那它就可称为是合理的。这的确是一个重大的改善。工具合理性将不再能独霸合理性领域;还有目标的实质合理性与之相抗衡。不过,即便有了这一修正,合理性在很大程度上依然是工具性的。① 在超越广泛的工具结构上,正是走向决策价值这一步是决定性的。

哲学启发法

科学家和哲学家们在建构理论中所作的是理性思考:他们构建出各种智力问题(intellectual problem),提出对这些问题的各种可能解法,检验且评估这些解法等。这种思考可能是工具性的,但它首先并不是指向信念的,甚至都不是对何为该问题的最佳解法的一种信念。因此我们前面讨论过的那些正反理由就没有处于最前台。即便如此,关于合理性的一种理论还是应当 164

① 我认为,这是 Nicholas Rescher 的观念:"合理性在于聪明地追求恰当的目标,"Rescher, *Rationality* (Oxford: Clarendon Press, 1988), p. vii。

能够涵括和阐明此类理论性活动。

哲学家们讨论推理问题时，往往集中于一个极为狭窄范围的思考，并视之为惟一具有合法性的推理模式。（有时候合理性的讨论者主要将合理性作为一种排斥工具，其主要目的在于把某些东西贬为不合理的，而不是正面地标出一种有效率且有效的装置。）若哲学家与科学家们在开创自己的理论时，其最好的推理方式都不能契合于他们标示为合法的那个狭小区域的话，那么，对何为合理思考的刻画显然便需要拓展了。

想出一种新的理论可以被看作是解决一个智力问题，或许是第一次注意到或者发现这个问题之后。那么，我们要去探究的就是这样两个相互关联的领域：问题的设定和问题的解决。设定问题的研究探问的是，如何发现和明确富有成效的问题，何种特征刻画了问题的情境，何种因素设定或者塑造了这些特征。解决问题的研究，看的是个人在该问题情境*之内*所做的与所完成的东西：她试图解决该问题或转换该问题的模式，所遇到的进展和拦路石，终极结果在她眼里是推进该问题还是未达到其目标的那种方式。（我们还可以研究，在引入她自己之外的一些标准与知识后，她在何种程度上确实推进或解决了该问题。）

什么是一个智力问题呢？问题的一般结构所得到的有用表征，是由人工智能和形式启发法(formal heuristics)的文献所提供的。在一个*界定良好的问题*之中，下述每一特征都得到了明确的规定与限定：

1. 一个*目的*，用来判断结果与状态的一个评价性准则。

2. 一个*初始状态*，由一种(起始)情境和可资利用的资源构成。

3. *可允许的操作*，可用来转换状态和资源的。这些可允许的操作是以规则形式陈述的，其规则可以用来转变那种初始状

态，然后再一次次地转换随致的各种转换状态。

4. **约束**，对过程中可以经历什么样的中间状态、可以达致什么样的最终状态、可以做什么样的操作（何时做、做多少次、以何种次序）等的约束。

5. 一个**结果**，亦即一种最终状态。

那么问题的一种**解法**也就是一系列可允许的操作，它们把初始状态转换为一个符合目标的结果，且在过程中的任何时刻都不违背任何约束。[①] 界定良好的问题的范例是诸如“传教士与食人族”（今天我们可以找到其他的描述方式）这样的形式谜题，以及各种形式证明的任务（“从公理开始，只用这些推理规则，证明这个定理”）。

165

即便人们所面对的绝大部分问题都无法这样精确地规定，但这个成分清单还是很有启发性的。[②] 人们经常并非只**面对**既

---

① Herbert Simon and Allen Newell, *Human Problem Solving* (Englewood Cliffs, N.J.: Prentice-Hall, 1972), pp. 71 - 105.

② 在这个问题模式中，目标是一种评价准则，因而给定任何结果，它是否满足那个准则，对这一问题都存在着一个清楚、确定且可得到的答案。然而在生活中，一个结果是否以及在何种程度上满足了那个准则却不是很清楚的。因为目标也许是长期的，一个即时的结果可能多少有助于获得该目标。一种评价有可能是相对的，是由他人做出的，就像各种建筑设计参与一场角逐时一样。因此，代理项评价，亦即期间评价（surrogate evaluations, interim ones），都可以被用作最终评价的估计。而且相关目标可能不止一个，也许还具有排序或者权重，但优先性通常更为模糊；目标可能是无序的，目标清单也是不确定的。（在发展一个计划时，有人可能发现它可以采取一个方向来有效地达致某一目标，而这个目标是他事先未曾关注过的；由此，这也**变成了**该项目的一个目标。）目标也可以不是预定的；个人可以选择目标或者对所得到的目标加以修改。

各种资源的获取可能有难易之别，有些资源会比其他资源更易于获取和利用，有些则成本更高。个人也可能并不清楚他的起始状态。可许可的操作也可能无法像数学中可许可的证明手段那样，能构成一张界定良好的清单——即使是这里，不是还是有人能够构建出一种不符合已有规则的、**新的**可接受的证明手段吗？问题有时候在于设计、创造或构造出达致某一目标的手段，而不是运用已然存在（转下页）

定的问题;他们的任务是去**提出**一个问题,亦即在他们身处的不成熟情形下去**找到**一个问题。构建出一个界定良好的问题,让它把捉你在情境中(将)看重的那些方面,这个任务并不是本身要作为界定良好的问题来对待(或处理)的一个任务。尽管如此,即使我们愿意放松它们的特殊性来适用于真实的生活情境,即使我们认同智力活动的一个重要部分并不在于解决这些问题,而在于获得这些问题,这种问题模式的四种成分清单(目标、初始状态、可许可的操作和约束条件——结果是解决该问题的产物,而不是问题的一个成分)依然是富有启发性的。一个模型若要阐明智力活动的一个阶段,不一定要阐明它的每一个阶段。

166

问题模型描述了情境如何呈现给具有该问题的个人,我们可以用它来理解他解决问题的种种尝试。然而,他对自己的"问题空间"中的某些特征可能认识有误;他拥有的资源可能比知道的要多,受到的约束也比他认为的要少。不过,影响他如何行动的还是他自己对情境的看法。在回顾时,我们可以试图解释他对问题空间中的某一特定方面是何以出错的;如果他卡住了,他本人也可以核查一下:他自己是否在某个方面出了错。

一个智力问题是如何被**找到的**,是如何被注意到的,是如何

(**接上页**)的手段。[由于行为人在发明这些手段的过程中必定会用到其他的手段,那么这个过程可否说成是更好地界定了她所面对的问题呢?]有些转换手段可能代价更高,或者若重复过繁将有出错之嫌。个人甚至会怀疑这是否是一个可被许可的操作。个人在自己的生活中没有坚持自己所提出的主张,我们可以用此来反对这一主张吗?我们能否在数学中运用归谬论证呢?为了实现一个有价值的目标,个人可以撒个无害的(不明朗的或有害的)谎言吗?各种约束也不一定要界定得非常好,它们的边界也可能是不明晰的。想一想,有没有可能搞出一种新方法削弱或绕开一个约束条件为什么看起来成立的理由,从而完全避免该约束。约束可能是个程度问题,违背它多少有点困难或要付出点代价。我们最好是把约束设想为一个**斜坡**(gradient)。(向那个方向运动的**难度**有多大?运动或者停在原处需要耗费多少能量?以时间或资源的方式来计算的话,代价有多大?必须克服多大的阻力?)

分离出来的，又是如何构建的呢？一旦问题构建出来后，人们又是如何设法**解决**它的呢（或至少是种尝试解）？对于智力工作的**方法**，亦即关于智力工作的原则、经验规则（rule of thumbs）、基本原理，以及如何构造此种智力工作等，我们没有任何确切的理论。文献对于有助于达到结论或解法的工具，称之为**启发法**，并且将这些与各种算法（algorithms）区别开来，后者保证在有限步内得到一种解（如果存在的话）。启发法的正式文献（formal literature）是试图构建出能由计算机编程且机械地适用的一些规则。而我们的目的与此不同：我们是为了找到非机械的规则与原则，它们能有益于一个聪明人，即使他还需要某种理解力、智慧与技巧才能应用它们。

我在此的兴趣是考察，建构哲学和其他智力问题（为了有成效地构建与解决智力问题）的经验规则。但这里所讨论的框架（即问题结构与启发规则）对于理解（某些）过去的智力工作是如何进行的，亦可能会有所助益。将“问题结构”作为思想史（intellectual history）的一个主要组织焦点来加以讨论的话，我们会偏离此处的主要关注；因此，对于有兴趣的读者，我提供一个补充的注释。①

---

① 卡尔·波普尔是思想史问题模式的最重要的倡导者。波普尔在其后期作品中，把他的这种进路置于他的“三个世界”的框架中：物体对象的世界 I，意识状态的世界 II，思维的客观内容的世界 III，这里包括科学的和诗意的思想（波普尔说，是艺术作品）。波普尔说，第三世界的居民是问题情境，它由问题、问题的背景（也就是嵌入在问题结构中的语言和各种理论）还有可用在这个问题上的各种概念框架和理论所构成。此外，第三世界包含批判性论证、理论体系和讨论论证的各种状态。波普尔强调，科学史不应该只是各种理论史，还应该是各种问题情境史，和它们是如何经由解决这些问题的尝试所修正的历史——这些尝试就是各种理论。（参见 Karl Popper, “On the Theory of Objective Mind”, 载于他的 *Objective Knowledge* [Oxford: Oxford Univ. Press, 1972], pp. 153 - 90，特别是 p. 170。）波普尔说，历史性理解不是从对第二世界的思维过程的分析，而是从对第三世界的关系分（转下页）

卡尔·波普尔把智力问题的解决看成是如此推进的一个过

(接上页)析中增长的。思想史学者研究智力产品,它们的结构性特征、相容性和理论性关系,也把这些产品作为对问题情境的回应来加以研究。"情境分析"就是对问题情境的一种理想化重构,行为者发现,这种情境使得他的行动或理论是可合理地理解的(就此可完成而言),表明在他看来这种行动或理论如何对于该情境是恰当的。

这意味着是如那个人看这个**情境**那样对于情境是恰当的,还是如那个人看**恰当性**那样对于这个情境是恰当的呢?(或者两者都是?)第一种情况会让历史学者如这个人所看到的那样来描述这个情境,然后试图表明那个人的行动对于该情境是恰当的,无论那个人是否以那种方式来看待恰当性。在这样做时,历史学者可以引入其他的恰当性标准,她自已时代的或者她认为是正确的任何标准。有趣的问题是,思维者的回应在那个情境下是否是恰当的。但是,那个人试图做的就是(在他所理解的那种问题情境中)得出一个解,且这个解根据**他的**恰当性标准是恰当的,或根据其时他的学科中的标准(他所理解的),而不是根据后来的标准或我们的标准。在解释为什么伽利略没有接受开普勒的行星运动规律时,波普尔回答道,那时伽利略不接受开普勒法则是有道理的,相反以一种大胆的过度简化来工作(p. 173)。现在这点契合**波普尔的**方法论;但是仅当它契合于伽利略的(他在那时所遵循的)方法论时,我们才有一个伽利略为什么不接受那些法则的解释。

John Passmore 把哲学史的一种模式——他喜欢的那种——描述为**问题史**。这种模式试图理解一个哲学家所面临的问题,即他试图回答的问题,然后试图追踪作为解决这些问题和回答这些问题的理论建构步骤。(John Passmore, "The Idea of a History of Philosophy", *History and Theory* 4 [1964 - 1965]: 3 - 32.)哲学家们长期以来面临的问题有多固定呢?当提出问题的理由有所不同,即使这些理由是类似的,不同的问题还会足够类似,从而对一个问题所提出的回答也能算作对另一个问题的回答吗?要是**各种问题**是相同的,那么(隐含的)可能答案的范围也必定是相同的,或至少是很大程度上重合的吗?"为什么是这样?""为什么是这样而不是那样?"前者往往是后者一种不明晰的形式。当两个历史时期相对于不同的"那样"而问同样的"这样"时,这些问题还有足够类似性,其答案还可以彼此竞争或有启发性吗?"那点是如何可能的?"这个问题,是"给定这点成立,那点是如何可能的?"这个问题不明晰的形式。当两个历史时期在面对所排除的不同的"这点"而困惑相同的"那点"是否有可能时,他们是在追问相同的问题吗?他们还是在谈论相同的"这点"吗?当两个理论家担忧于自由意志问题时,一个是因为他把神圣先知看作为给定的,另一个则是因为他把普遍的因果决定论看作是既定的,他们是在探讨相同的问题或谈及相同的事情吗?即使这些问题不是恒定的,问题导向的历史还是能够研究:过去的学者是如何努力去解决**他们的**问题,为什么这些哲学家的问题长期以来一直在变化。

(转下页)

程，即检验临时解决方案并对这些方案加以批判，据此来修正

（接上页）　在一个众所周知的纲领中，英国的政治思想史学家 Skinner 提出了一个拒绝问题模式的探究大纲。他说，不要把政治学者看作是为长期问题提供一个答案，或看作是为永恒的主题提供各种立场，或甚至是在试图解决当下的**知识**问题。相反，他们的作品是介入具体的冲突。我们应该把他们的主要意图，他们的以言行事(illocutory)行为看作是在做一件事：在具体的社会与政治冲突中支持一方，为一个派别的立场做论证等。（参见 Skinner，"Meaning and Understanding in the History of Ideas", *History and Theory* 8[969]：3－53。）作者的意图是一个具体的意图，特属于那个具体场合。（Skinner 也允许研究其他的东西，但是他把介入具体争议的确定和研究看作他的思想史模式的核心。）

然而，对于每一个具体的冲突，成千上万的人都会采取各不相同的立场。我们对这些作家感兴趣的理由不在于他们在冲突中站了队，而是在于他们说了些有益的东西，实际上看来是说了超越了那些具体冲突并且能更一般地应用的东西。若不是这样，那么鉴定出这些作家（假定）必定进入的那些具体冲突，这个任务就不会如此艰难。实际上，一个作者被看作对何种冲突采取了一个立场，很可能取决于他写作的确切日期。不同的年份，有不同的冲突，从而有不同的介入。

当然，绝大部分时间里都有着这样或那样的社会和政治冲突在上演着。因此毫不奇怪，思想史学者能够找到作品所关注的一种冲突。若一个作者说的东西多少有点广泛的适用性，它就将对许多不同的冲突有影响。他在一个特定的时间说了的话，对当时流行的冲突具有影响，这并不意味着他的意图（或以言行事的行动）就是在那个冲突中站队，肯定不是说他的意图就**只**在于站队。因为这个作者的意图可能在于提出一个普遍性理论或道理，从而具有广泛的意义和适用性。他的以言行事的行动，如果我们需要引入这个术语的话，可能是**理论化**。这个政治学者也许是试图说出适用于（许多）其他语境和时期的永恒道理，这样一来，把他们的言说只看作是对一个具体的语境或冲突，这就扭曲了他们的旨意。

有些情形中，我们会赞同社会学家和历史学者的观点，作者的一个目的确实就是对一个冲突提出一个具体的原因或站队。不过，即便在这样的情形中，我们还是必须问：作者为什么要通过提出抽象的理论性内容，即普遍原则来这样做呢？要赢得他人的支持和巩固自己的立场，他不可能仅仅是在那里宣告他对那一方的偏爱；他必须弄出理由来说服他们。理由可能是特殊的，但它们也可能是一般的理论性考虑，可以很好地适用于大范围的情形，且在这个场景下支持一方。如果它们适用的其他情形是其他人早已接受的，那么（通过一般性推理）这些其他的情形将援引来作为证据，对当下的相关情形所做的判断提供支持。

这样，即使一个作者确实意图介入一场具体的冲突，即使他的主旨不在于**理论化**，我们还是会对他的作品感兴趣，原因不是他介入了一边，而是他成功地提出了一个一般性的且可能是有说服力的理论，这个理论适用于大范围的情形和历（转下页）

原问题，提出新的临时解决方案，以此类推，直至问题被解决。

（接上页）史情境等。我们感兴趣的程度将匹配于他成功地提出的理论所具有的适用性、吸引力的和说服力。（记住，有成千上万的人都支持这边或那边，但我们并不会去详细地研究他们。）我们对理论者感兴趣的东西，亦即使得他们重要的东西，并不是他们也站队了——*如果*他确实站了的话——而是他所发展的理论。即使这个作者并不仅仅是想理论化，而是试图通过抽象和一般的推理来证成，若不聚焦于*他*所聚焦的东西，不聚集于他的一般性立场的支持性理由的结构，当它影响到这个立场的恰当性和可接受性时，我们就不可能理解他所做的事情，情况还是如此。若作者以言行事的行为是有道理的，我们的主要关注将是探讨他是否*确实*有道理，并且是在何种程度上。那么思想史必定在很大程度上是思想、理论和有理据的立场的历史，而不是一个权力游戏中可鉴定的智力运动史。（在另一篇论文中，Skinner 确实注意到：即使一个理论者是愤世嫉俗的，他的公共证成性理由将约束他能采纳的立场和做的事情。参见 Quentin Skinner, "Some Problems in the Analysis of Political Thought and Action"，载于 *Meaning and Context: Quentin Skinner and His Critics*, ed. James Tully [Princeton: Princeton Univ. Press, 1988], pp. 110 - 114。）由此，我们又回到了智力问题领域，并且试图解决或促进它们。

思想史学者用了广泛的要素来理解问题的设定和塑造，对此做一个一般性的分类是很有用的。Peter Gay 在 *Art and Act*(New York: Harper and Row, 1976)一书中列出了三种类型：

1. *文化*：社会和经济要素，社会的需要和问题，宗教和政治压力，通常是制度性的。

2. *工艺*：一个主题或学科的技艺、传统和工具。我们可以使用 Thomas Kuhn 的术语，称之为学科基质(disciplinary matrix)：属于这个学科里广为人知的和可用的那些工具、技艺、继承的问题、知识体系和讨论的当下状态，还有预计参与者会应用的那些标准和评价准则。

3. *私人领域*：个人的家庭、内部的心理生活、焦虑、幻想、捍卫和无意识的需要，还有更窄意义上的*传记*。

我们还可以在这些要素上再加上两类：

4. 个体对于判断一种理论和侦察出一个问题的*个人性智力标准*。（例如，爱因斯坦就认为，引力质量和惯性质量的等价是需要解释的东西。对称性在没理由出现的地方出现了，在对称性应该占据的地方出现了不对称性——这些因素和类似的要素，在审美上设立边界，可能会为哲学设定一个要沉思的问题。）这些个人性标准不一定在学科内是广泛的，尽管它们可以*变得*如此，如果遵循它们能够获得强有力的理论的话，这就会使得这些标准在他人看来是显著的。

5. 社会中的*普遍思维模式*，并不必然是立足于制度的。这包括一种信念框架，诸如 P. F. Strawson 的描述形而上学；一种普遍的因果和解释性原则框架；（转下页）

只有在极罕见的情形中，我们才会得到详尽的记事本，从中见到思想者如何规划并发展自己的思考。① 我们通常所见的都是成品，而这些成品通常由陈述的问题(可能并非原来的那个问

(接上页)何种事物需要解释，何种事物不需要解释的标记；何类要素可诉诸作为解释性要素，何种要素可作为一种理论的证据的标记。

给定对特定问题情境成分的一种规定(它的目标、初始状态和资源、可许可的操作以及各种约束)，我们可以继续探究这五种要素中的哪些要素塑造了这些特定的成分。我们能够形成该影响的各种可能性矩阵，对于一个具体的问题，我们可以探究每一列如何塑造了每一行(例如，学科基质如何固定或塑造了各种约束，文化如何塑造了目标等。)这不是一种设定问题的**理论**；它是对各种各样影响的分类，在这种结构内可以组织各种历史化探究，即一种可以追问的问题的核查清单。我们可以追问：什么东西使得**这些对他而言**是目标、初始状态和资源、可许可的操作和各种约束呢？那个人如何对他的情境结构化，并且认为他本身正在面临特殊的问题，无论其成分是多么的混杂和凌乱不清？

学科史着重于学科基质如何影响问题情境，因此影响随致的智力产物。更宽泛的历史会查看所有五种因素。但是由于智力产物的制造者通常相关于前期产物来定位他们的工作，批评、修正或发展它们，这样来区分出他们的新作品，因此，思想史的指导性论点是学科基质将起很大的作用。思想史学者的任务并不终止于研究一种理论或思想的创造，她也研究它是如何扩散以及它在学科内和更广泛的社会里具有的影响，包括它对于基质里的五种要素的影响(文化、工艺等)。什么东西有助于为新思想提供空间，且以致它甚至被认为是可能的呢？(参见 Hans Blumenberg, *The Legitimacy of the Modern Age* [Cambridge, Mass.: M. I. T. Press, 1983], pp. 457 - 481。)什么东西决定一种新思想得到多少注意，谁在本学科内、在其他学科内，在整个社会来协助传播，什么样的社会和个人激励会导致这样做呢？谁选择把麦克风放在某些思想前，为什么他们选择来放大它们呢？(参见 Bruno Latour, *Science in Action* [Cambridge, Mass.: Harvard Univ. Press, 1987]，讨论在科学中联盟网络的形成过程。)一种思想在传播中是如何得到修正和淡化的？思想史学者也可探究一种思想与学科内及学科外的思想的竞争中表现如何。特别是，存在着那种合理且客观的标准，据之竞争中的胜利者要优于失败者吗？即便存在着客观的学科标准标明一个竞争者优于所有其他的竞争者，标准的可能范围大也意味着我们必须探究为什么当时援引了那些具体的标准。

① 即便这些也可能是经其编辑与“净化”过了的。米开朗琪罗(Michelangelo)与他留下来的信件与画作就是如此，那是一本设计出来以支撑他的这一看法的文集：他自己完全是无师自通的，且在自己的计划中无一例外都获得了成功。

题)、对该问题提出的解法或许还有对此解法的有些可能批评的回应所组成。从这一最终产品来重构出得到它的整个过程,也许还是由许多重复的步骤所组成的,这绝非易事。启发式原则会进入每一个阶段。我们可以试着确定:构建和选择问题的原则或规则,构建问题的临时解的原则,批评问题的可能解法的原则,还有在面对对早前提出的解法的批评与困难时,重构和修正问题的原则。但波普尔对这些阶段的概述,提醒着我们留意那些不同种类的启发式程序,它们或许渗入了对一个智力成果的建构之中。

167

当前讨论的一个问题是:在多大的程度上,智力产品是由运用于大范围领域和主题上的**一般**启发法所导致的;在多大程度上,它们是由特定的启发法所导致的,这种启发法体现了大量有关某一特定智力领域(或子域)的结构、模式与程序方面的信息。[①] 正如我们可以(在部分程度上)通过特定思想家们所偏爱的启发法类型来刻画他们的智力**风格**,我们也可以通过各种主题所运用的是一般的还是特殊的启发法类型来刻画这些主题。

这里有些关于启发法的原则与程序的例子,如摸彩例子,它们几乎全是指向对一个智力问题产生第一种临时解法的最早期阶段。(我这里想的全是**理论性**问题;其他类型的启发法将应用于其他类型的智力问题。)这些**特殊的**经验规则本身就是有益的。我们还应当留意的是,合理性领域比评价正反证据还要宽泛得多。请记住,这些只是启发性规则——无法保证它们在任何具体的应用中会成功。

---

① 参见 Pat Langley, Herbert Simon, Gary Bradshaw, and Jan Zytkin, *Scientific Discovery* (Cambridge, Mass.: M. I. T. 1987), pp. 3-46, 49-59; D. N. Perkins and Gavriel Salomon, "Are Cognitive Skills Context-Bound?" *Educational Researcher* 18, no. 1 (January-February 1989): 16-25。

1. 当多种思想立场(intellectual positions)之间的冲突长期存在,且无任何解决之道或者大的进展时,去寻求所有竞争立场所共有的假设或预设。① 试着否定这一假设,并在这样开放出来的新空间内,尝试建构一个新的立场。

前面缺乏智力解法的一种可能解释是,所有竞争立场或竞争方都受困于这样一种思想框架,它排除了对该问题有一个恰当解。(另一种可能性是,目前的思想框架是适当的,但人们不够聪明而无法在此框架内解决该问题。)在解法被构建出之后,曾被所有人视为理所当然的那个假设就会显得有点武断了。

但是个人如何能鉴定一个为所有人(包括他自己)视为理所当然的假设呢?试想这样一个问题:如何用四条直线连接起九个点,且笔不离纸。② 个人如何鉴定这个假设就是排除了解法的那个假设来呢?或许通过将每一件事都说得明确无比——"每条所画的线都必须停在一个点上"——然后,去察看这一明确的陈述是否确实在问题的原初条件之中。[或者对于其他的问题,这样说(即"记住,你允许画到点外去")也是有益的吗?]

168

现在这里有一个有关假设的例子,在我看来这个假设构造了一个问题。就爱波罗悖论(Einstein-Podolsky-Rosen Paradox)(EPR 悖论)和贝尔不等式而言,当说分离粒子之间的联系违反了局域性(locality)时,它假定了空间的拓扑结构(topology)与度量标准(metric)是不变的。如果两个粒子的生

---

① 参见 Frank Ramsey, *The Foundations of Mathematics and Other Logical Essays* (London: Routledge and Kegan Paul, 1931), pp. 115 - 116。

②

成使空间的拓扑结构与度量标准发生了改变的话，例如，若两个粒子分离时在二者之间形成了一个不断伸长的虫洞(wormhole)的话，那么(在这个有新的度量标准的空间来看)这两个粒子之间的作用就有可能是完全局域性的了。通过抛弃这一拓扑结构与度量标准确定不变的假设，我们就可以在微分几何的范围内探究一种替代性理论，它能产生一种适当的可变拓扑结构，并寻找与查证其可检验的结果。

2. 我相信，往往只有在考虑一种新的极端可能性之后，我们才能明确地识别出根本假设。(因此得到新可能性的通常步骤就不是，第一，识别出那一根本假设；第二，否定它或假定它是不成立的；第三，再看会蹦出什么样的新可能性。)然而，我们还是可以利用这样一种全新的可能性(一旦它被想出来以后)来追问，它所违背的根本假设是什么，还有哪些其他的可能性也是违背这一假设的，如果抛弃这一假设，那么还有什么样的新框架是适当的等。

3. 特别要注意那些意料之外的对称性或不对称性，这些都是没有任何特殊的理由认为它们应该成立，或有某些理由认为它们应该不成立的。[爱因斯坦在论狭义相对论的论文开篇即说，麦克斯韦(Maxwell)的电动力学，按通常的理解，会导致“看来并非该现象固有的一些不对称性”。①]如果一种属性中存在着不对称性，然而一个语境中的所有相关因素都对该属性产生的是对称性的影响，那么就去考察一个更宽泛的语境，以找到一个具有不对称影响的相关因素。

① 对爱因斯坦思想中的对称性与不对称性的讨论，参见 Gerald Holton, “On Trying to understand Scientific Genius”, 重刊于他的 *Thematic Origins of Scientific Thought* (Cambridge, Mass. : Harvard Univ. Press, 1973), pp. 353 - 380。

4. 把已取得丰硕成果的一项操作或过程运用到新情形中，把新情形与旧情形之中的差别做出恰当修正之后，它们在恰当方面是类似的。[例如，考虑波斯特(Emil Post)用来生成逻辑学中的系列定理的生成观，把它应用于语言学中，其目标就变成了生成一种语言中的语法系统——乔姆斯基(Noam Chomsky)若用此法的话，他本可以达到他对语言学的初始重构的。]这只是一个更一般准则(maxim)的例示：

5. 尝试用其他发达领域的模型或类比，把你正处理的混乱材料进行结构化。

但我们如何能找到一个**富有成效**的类比，帮助我们解决正试图构造的那个**对象域**(target area)之中的问题呢？假设你已经有了一个问题，它是用所述标准结构(目标、初始状态与资源，经许可的操作和约束条件)而陈述出的——一个有关讨论归纳的作品在此给出三条建议：①

A. 从该对象的初始状态开始，系统地修正这一状态，直至达到你事先要求(command)的那种知识结构。用该结构中的可许可操作，为你的对象体系建构相应的可许可操作，并将它们应用到初始状态，从而查看是否达致目标。(如果它们能让你接近目标，思考如何调整或修正这些对应的操作以使它们能精准地达到该目标。)

B. 从对象中的目标 G 开始，在其他领域中寻找类似的目标 G'。在该领域中寻找达到目标 G' 的操作 O'。在对象域中建构起对应的各种操作 O。检查这些操作在对象域中是否能产生出

---

① John Holland, Keith Holyoak, Richard Nisbett, and Pual Thagard, *Induction: Processes of Inference, Learning and Discovery* (Cambridge, Mass.: M. I. T. Press, 1986), pp. 286 - 319.

目标 G。

C. 从对象的目标 G 和对象的初始状态开始。构建出一种直觉，判断初始状态中哪些特征 $F1$，……，Fn 与达到目标是相关的，也就是说，它们最终将进入达成目标的那些操作当中去。然后，只用这些特征来构建对象体系中的初始状态的描述 D。找到另外一个具有相似结构特征 $F1'$，$F2'$，…，Fn' 的领域。将对象的目标 G 转换成该领域中的相应目标 G'。在那一领域中看何种操作 O' 从（带有特征 $F1'$，$F2'$，…，Fn' 的）初始状态得到目标 G' 的过程。把该操作 O' 转换成对象域中的对应操作 O。检查 O 在对象域中是否可以产生目标 G。

这些*特定程序*是为了形成一种可能富有成效的类比物，我们也许会用它们来重构一个思想家会经历的（也许已经经历的）思想过程。其他程序并不那么特定，为了历史学家试图重构该智力产品过去的实际生产过程而提出问题。比方说，你沉浸于某个问题，然后神游、探索，且留神那些暗示着富有成效类比物的线索。（遵照这种建议，个人就会遍览书架上的书名、一页页地乱翻书、随便乱逛等）这里面的假设是，当你沉浸于一个问题时，外在的帮助能让你灵光一闪，找到富有成果的类比物。要注意，正如机会只给有准备者那样，经历上述三个明确步骤也有助于更易找到同类物。即便这些步骤失败了，你也能更清楚我们所需要的是何种类比物，由此（某种程度上是不自觉地）有助于指引你的寻找。

170

当个人可援引的知识结构与理论贮备越大时，他就能越好地找到对问题有成效的类比物。（当面对一个难题，他人的各种尝试均告失败时，含有各种结构和已解决问题的那种“工具包”，就必得涵括极为不同领域里的例子。若其他人将来还是使用差不多的工具包，他们就还是不会成功。新进路的建构者往往具

有获得了不同寻常的各形各色资源的历史。①)大多数思想家往往会放弃自己早先所拥有的智力资本;在某种意义上讲,增加那些新结构与新工具是可欲的,它们除了直接的用途以外,还能用作同类物资源且刺激结构性想象。

6. 从目标向后推,从初始状态向前推,看你能否使这一洲际铁路贯通起来。②

7. 把一个困难的问题还原为一系列相对简易的问题,然后试用其他启发法来解决这些问题。③

8. 考察**极端**情形,考虑若把某些参数设置为零或无限值,看看会出现何种结果,然后参照这种极端行为来重新考量居间情形。

9. 考察并列举出问题的正确答案所必备的一般性特征。然后寻找具备这些特征的东西。如果你找到了与一种特征完全相符的东西,再也找不到任何其他的特征,那就试着证明它是惟一满足条件的解法,因此也就是该问题的解法。如果你找不到任何符合条件的,那就试着证明不存在满足全部条件的;如果成功了,那我们就得到了一个不可能性解,就像在社会选择理论或不确定条件下的决策理论中有时所见到的那样。④ 简单地去掉

---

① 关于这一点,参见 Howard Gardner, *The Creators of the Modern Era*(即出)。

② Simon and Newell, *Human Problem Solving*.

③ 参见 Georg Polya, *Patterns of Plausible Inference*, 2d. ed. (Princeton: Princeton Univ. Press, 1986)。

④ 参见 Keneth Arrow, *Social Choice and Individual Values* (New York: John Wiley, 1951); Amartya Sen, "Social Choice Theory", 载于 *Handbook of Mathematical Economics*, ed. K. J. Arrow and M. Intriligator (Armsterdam: North Holland, 1985); John Milnor, "Games against Nature", 载于 *Decision Processes*, ed. R. M. Thrall, C. H. Coombs, and R. L. Davis (New York: John Wiley, 1954), pp. 49-60; Luce and Raiffa, *Games and Decisions*, pp. 286-298。

一个条件可能会恢复一致性,但却有太多的东西太容易满足余下的条件。而稍微放松一个条件,将在恢复一致性的同时仍留有一个严格的条件集,使只有惟一的一个对象能够满足它们,这是一个可欲的结果。任务就是考虑对解法所提出的条件中哪个应当被修正,或去掉哪个条件且随后加上什么东西取代它,然后我们再去考察是否有能满足这一新构建的条件集的东西。

171　10. 对于一个新的具体想法,可以构建出一点形式结构或模式,把它嵌入其中,然后再探讨它的属性与含义。①

11. 对一过程、概念或现象找到更为**抽象**的一种描述,并探究其属性以得到一种更一般且强有力的结果;只要得出的结果还是强有力的,那么就不断提高该描述的抽象性。

12. 在探究某种关系 R(如解释或证成)时,也要考虑到由 R 所归纳出的整个领域的结构。这个全域结构引起了什么样的特殊问题? 对 R 作何种调整能生成一个不同但更好的全域结构呢?

13. 转变已知的现象来发现新现象。② 当已知 a 与 b 处于关系 R 时:(1) 探查现象的范围:即可替代 a 并且仍与这个或那个东西处于关系 R 的事物清单是什么? 可替代 b 并且仍处于关系 R 的事物清单是什么?(2) 通过描述划定每个范围的属性来刻画现象的内容。(3) 若用其他某种显然类似于关系 R 的 R' 来替代它,探究事情会如何改变?(4) 探究在一个具体的过

---

① 一个非常温和的例子,参见我的 *Anarchy, State and Utopia*, pp,59 - 64,以及 *Philosophical Explanation*, pp. 363 - 380, 388 - 390 关于报复性惩罚的 $r \times H$ 结构的讨论。要旨在于,即便是这样一个极为简易的结构,都能产生出有趣的结果来。另一个温和模型的例子是 *Anarchy, State and Utopia* 中的正义的资格理论,也是类比于形式体系的一般结构(其中有公理、推导规则以及推导出的定理)而建立起来的。

② 参见 Langley, Simon, Bradshaw, and Zytkin, *Scientific Discovery*。

程中,用一个成分代替另一个成分会导致怎样的新现象?

14. 如果你试图通过提出一个决策或描述是一清二楚的情形,从而把这种描述或决策施加给一个与之并列但不那么清楚的情形,那么,陈述出使得这两种情形在明晰性出现不同的那种差别,表明这种差别的存在为什么并不使得我们可恰当地把两种情形判别为有差别的或甚至相反的。[①]

15. 最近出现了一种新的程序来生成问题、凸显困境和刺激细节性想法:建立现象或程序的一种计算机模拟。

16. 为科学和哲学领域中产生富有成效的**思想试验**,构建出一些原则,这都是有益处的。(回想一下奎因论人类学家作彻底翻译[radical translation]、普特南论孪生地球、维特根斯坦论建筑师和体验机[experience machine]的例子。)[②]

在我看来,上述这些特定的启发原则都值得一试。我在这里描述它们,也是希望能激发更多的人——不仅限于哲学家——在智力探究的不同阶段上建构出富有成效的启发性原则。不是"机械"原则——这种原则的应用需要涉及大量的知识以及对材料的一种"感觉"。让人有点惊异的是,在训练哲学学生时,老师们没有做出任何这种严肃的努力,以构建出这种智力探索的经验法则。学生所看到的是那些成品(书和文章)和其老师对这些作品所作的考察,然后让学生自己去琢磨这些东西是怎么弄出来的。一些明晰的提示也许就是有用的。 172

---

① 参见我的"Newcomb's Problem and Two Principles of Choice",载于 *Essays in Honor of C. G. Hempel*, ed. N. Rescher et al. (Dordrecht: Reidel, 1969), pp. 135 – 136。

② 近期对科学中的思想试验的讨论,请见 Nancy Nersessian, "How Do Scientists Think?" 载于 *Cognitive Models of Science*, ed. Ronald Giere (Minneapolis: University of Minnesota Press, 1992), 特别 pp. 25 – 35, and David Gooding, "The Procedural Turn", in ibid., 特别是 pp. 69 – 72。

合理性的想象

合理的理论思维有着诸多不同的工具性程序与模式。其目的在于事物而不是信念——如新颖和富有成效的智力产品——因此它们并不总是聚焦于正反理由的挑选或评价(各种启发性的原则即是例子)。同样,信念的合理性也不仅仅是一个应用(机械的)规则来权衡既定理由的问题。想象在其中起着重要作用。我这里的想象意指的是想出各种新的和富有成效的可能性的那种能力。那么,启发性规则的清单也许是一个理论的开始,规定这种想象是什么及其如何起作用。

决定陈述可信度的,是它的正反理由,以及削弱和巩固这些理由的其他陈述。然而,侦查出还有其他可能性削弱陈述可信度,这并不是一个机械问题。证据是提出来支持一个假说的,但是所有相关变量都得到控制了吗?每个相关变量都标示着一种也许可用来解释该数据的替代假说。这也就是为什么每种相关变量都必须可控,否则,现有数据给该假说所提供的支持就不再那么有力——尽管人们需要想象与聪明才智,才能探查出何种未经考虑的变量可能对该数据的生成是有作用的。在应用信念的规则 1 时,想象也会进来。还存在其他相竞争的、不相容的陈述有更高的可信值吗?构建出最值得考虑的替代陈述不是一个机械的问题——相对论是牛顿力学的一种替代,但唯有爱因斯坦才成功地把它构建出来了——有时甚至是搞清楚它是一种替代,即一种相竞争和不相容的陈述,也不是个机械的问题。

新选项的诞生无论是在信念上还是在行动上都有重要的作用。行动是在各种选项之中选择的。在现有选项中做出更好的选择乃是改善结果的一种方式。另一种方式则是扩大选项范围

以涵括有希望的新选项。富有想象力地构建出一种至今尚未得到考虑的新选项,这也许有可能做出最大的改善。也许存在有益的规则决定何时来寻找这样的新选项,但结果仍取决于实际上发现它们。这里没有任何机械的(或算法的)程序来生成最有希望的选项——无论如何,我们知道的那些启发性准则,没有一个能帮上忙。 173

这些考虑能否使想象成为合理性的一个组成部分呢?有些人或许坚称,合理性只在于在那些**既定**选项(无论是行动还是信念)中选择最好的。这看来是一种既不必要也是专断的切割。想象即便不是合理性的一个组成部分,也是合理性的一个协作者,是实现合理性目标的一个重要手段。在某些情境下,生成新的可能性并大略地选择,这比只在既有选项中精挑细选要更有收获。新选项中的次优选择很可能远远好于旧选项中的最优选项。因此,培养我们对相关问题的想象能力与增强我们的分辨能力乃是同等重要的。

不去探索及检验其他可想象的可能性,而只拘泥于**既定**选项,这种合理性程序乃是鼠目寸光的。即使此种程序为我们做得很好,它们也仍然将我们限在一种局域最优上。在讨论极大值问题的文献中,所使用的著名类比就是关注个人在一个地域去攀登最高点。假如有一个近视眼者,只能看到十尺远,可能会遵循这样的程序:扫视四周,然后登上他所能见的(十尺以内的)最高点;一次次重复此程序,直到他的视野中再也没有比他现在所矗立之地更高的点了,才停下来。那么,如果这个人是从一座山的山坡处开始攀登的,则这一过程会将他带到山顶;但是,却绝不会把他带到另一座更高的山顶。这个过程会把他带到一个局域的最高点,亦即一种局域最优,但却无法达致一个总体的最高点,亦即一种全域最优。

如果不去充满想象地生成并检验各种新的可能性,则合理性本身将只能带领我们达致局域最优,亦即既有选项中的最优。合理性能帮我们达到这一步,已经是它的一个很大的优点了。但是我们仍需要谨慎,以确保合理性不会把我们束缚于此。合理性不仅十分容易而且倾向于变成这样的装置,把新可能性的虚拟生成和检验视为不合理的,从而予以排除。探究新可能性的过程会是不完美的,显然也是会有所浪费的;所探讨的许多可能性会表明是无用的。然而合理性也必须宽容此点,不要事先就要求成功的保证。

发现和证成(想出一些假说且评估这些假说的可信性)的语境不能被彻底分离。因为为了评估某一假说的可信性,我们必须想出并且考察与该假说不相容的最好的那种选项。(我们能174 够明确地相对于一个给定的不相容选项来评估一种假说的可信性,但是这无法向我们给予后者任何超然的结论。)

"还有什么新的可能性",这个问题是人类进步的第一步,即在制作、活动、合作、思考和生活等方面生成新理论、新发明和新方式。提出这个问题就要有打破传统的意愿,即冒险进入未知领域。回答则要求具有想出新的且富有成效的可能性的那种能力;也就是说,它要求想象力。

并不是每个人都想探讨所有领域的各种可能性,每个人都这样做也是效率低下的。我们受益于他人的活动,也受益于我们之间的差别。正是这种差别导致别人去做、去想我们自己不会去做、也不会去想的事情。科学哲学家一直试图构建出机械程序,它们能导致科学家对于是否接受某个特定的科学理论做出一模一样的决定(在那个情境下,能够得到具体的证据,已构建出了各种具体的不同理论等)。但库恩(Thomas Kuhn)注意到,看法的差别在科学继续的过程中有着重大的作用。有些科

学家探讨和提出新理论，而其他科学家同样有力地捍卫和修正既有理论来处理新现象，这时事情会进展良好。正是这些不同路径上的发展才最终产生了对不同理论的能力与局限性的详尽认识，由此引出科学家表明他们具有何种普遍的一致性。①

哈耶克(Frederick Hayek)强调，在社会生活中，我们是如何得益于这种个体的探讨的，他们尝试新的生产方法，发展新产品，试验新的行为模式和不同的生活方式。我们受惠于以这种方式来探讨的一般自由，即使我们自己并不利用这种自由。他们的探究得到一些成就后，其他人也会仿效他们，最终我们也会这样做。即使我们不做，我们也能从其他人这样做的活动中得益，因此也从原始探险者和革新者的活动中得益。也许这种探险和革新中的绝大部分都是徒劳无功的，但是成本主要由那些选择探险的人所承担了；当(相对少的)几个有成效的革新扩展其效果时，我们都会受益。② 我们的经济生活、智力生活和政治生活的社会性质，使我们能够受益于我们自己并不具有的那种想象力——没有人能在所有领域都有同样的想象力，只要这要求我们受限于其中的那种敏锐的注意力。③

合理性有两面性。合理性令人兴奋的是它的锋利的切割边和胆大包天。有精确的标准、充分考虑反面理由和削弱因素，这些使得规则 1 是个强有力的武器：不要相信这样的陈述，如果与它不相容的陈述更可信的话。苏格拉底在刺透华而不实的信 175

---

① 参见 Thomas Kuhn, "Objectivity, Value Judgment, and Theory Choice", 载于 Kuhn, *The Essential Tension* (Chicago: Univ. of Chicago Press, 1977), pp. 331 - 332。

② F. A. Hayek, *The Constitution of Liberty* (Chicago: Univ. of Chicago Press, 1960), ch. 2.

③ 对于警惕性的有限，参见我的 *The Examined Life*, pp. 40 - 42。正是 Hayek 把文明化的程度界定为我们能从我们并不具有的知识中得益的程度。

念时的那种一丝不苟和勇气，现在的年轻人正如适逢其时的年轻听众一样为之而激动。合理性的胆大包天在于它愿意构建出以前甚至不在考虑之列的东西，支持以前由于让人害怕而被否定掉的信念——只要考虑一切理由后，这些确实是比其他竞争者更可信。就决策而言，合理性也毫不尊重专断的约束：问题在于哪个行动确实最大化了一切相关的功能，至于那个行动以前闻所未闻，这一点也不要紧。在科学探究里所冒出来那种不可思议的、耀眼夺目的、令人震惊的，有时是让人不安的理论中，合理性的荣耀得到了最清楚的体现，但在更为平常的语境下也有所体现。合理性的第一面是浪漫，卡尔·波普尔对科学有个令人激动的描述，很好体现了这点。他称科学为一种尖锐批评，一种对大胆前卫理论的测试，经常在分歧的深渊之上走钢丝绳。即使我们知道这个故事过于简化，有所欠缺，但它依然是激动人心的。

但是合理性的能力并不仅仅在于它耀眼夺目的个体成功。合理性还有累积性力量。一个既定的合理决策可能比不那么合理的决策不会好很多，但是它产生的决策情境却不同于另一个决策会产生的那种情境；在这种新的决策情境下，一个进一步的合理决策会产生其影响，而这继续会导致另一个决策情境。长此以往，合理性上的细小差别复合起来，产生非常不同的结果。一个既定的信念可能不会比一个不相容的信念可信很多，但是这两个信念会支持不同的进一步的陈述，随着这种路径上的可信度差别的不断复合，这会导致极为不同的信念体系。在下棋时，有的棋手会通过小优势的累积而摧毁对手；其他的棋手做出大胆的且锐利的攻击与牺牲。合理性就像最伟大的国际象棋冠军那样，两者都干。

我们的探讨让我们得到了新的合理性原则。一个合理决策

的原则要求决策价值的最大化，这让我们超越了合理性的简单工具结构。两个原则管辖着合理的(甚至显然是纯粹理论的)信念，消解了理论与实践二元性：不要相信比不相容陈述可信值低的陈述(理论成分)，但然后仅当相信一个陈述的期望效用比不相信它更高时才相信(实践成分)。信念的合理性涉及两方面：得到使信念可信的理由的支持，且由可靠地产生真信念的过程产生。我们提出的理由的进化性论说，解释了这两个方面中令人困惑的联系，但颠倒了康德的“哥白尼式革命”的方向。 176

传统典型地认为合理性最重要的作用是确定人性的特殊性，而进化的视角对于合理性的本质与地位也产生了一个全新的画面。合理性是对划定的目的和功能的进化性适应。它得到挑选，也设计来配合人类进化时期长期成立的事实而起作用，无论人类是否能够合理地证实这些事实。许多哲学传统上棘手的(即无法合理地解决的)问题，可能是源自把合理性扩展出这种有限功能的那种企图。这些问题包括归纳问题、他心问题、外部世界问题和证成目的的问题——康德试图把有原则的行为作为行为的唯一终极标准，这是合理性超越其边界的另一种扩展。我们的合理性是出于其他有限目的而塑造出现的，有些条件恰恰是为了与合理性配套起作用而与之并肩进化而来的，若说合理性自身能够证实(demonstrate)所有这些条件为真，这即使不是不可能，也会是中六合彩。

我们已经探讨了工具合理性的局限性，但这些不应该被过分强调。工具合理性是一个强有力的训练工具，其他的每一种合理性观念要想完整的话，都要把它作为一个重要的部分包括和灌输进来的。我们对于目标合理性提出了诸种条件，由此得出的观念要比工具合理性更为宽泛，后者只是要有效和有效率地追求合理的目标。然而，这些条件还不能完全确定这些目标

或欲望的实质合理性，这点也许是我们应该庆幸的。一个完全确定的实质合理性理论开启了由外部强加的、专横要求之门。缺乏这样一种理论的话，的确某些可反驳的欲望，包括一些不道德的欲望，会得到允许，但这正是一种实质的伦理学理论要对付的——尽管哲学家持之以恒地想把伦理学归入合理性之中。工具合理性为我们留下了自主地追求**自己的**目标的空间。

不仅如此，我们已经发展了一种合理性概念，它甚至超越了宽化的工具合理性概念（作为对合理目标的有效追求），从而包含了象征的和证据性的合理性观念。无疑，正如因果工具性那样，我们对象征性要素与证据性要素的考虑也可以具有进化的起源。无论象征能力的进化功能是什么——无论是加强其他的欲望，在贫困时期通过加强它们的实际对象而维持它们，还是使得人们在囚徒困境中进行协作，否则本来是不会发生合作的——这个功能都不一定是我们当前的目标，更不用说最大化全面适合度了。一旦我们具有了这种能力，无论其起因是什么，我们就可以出于我们自己的理由和目的来使用它。就像我们的数学能力一样，我们同样不应该把这种能力只限于服务其初始功能。

177

合理性具有进化基础，但这并不注定我们要按前面标示的那种进化轨迹来继续。（但是，知道一个特质的进化功能和认识到它不再服务该功能，这也并不保证我们将选择改变这种特质，即使我们能够这样做。我们可以保持该特质，因为它是一种我们现在独立地重视的生活模式的推动力，因此我们赋予它一种新的功能。）我们已经使用我们的合理性能力来识知这种进化基础，尽管这些能力并不是为了那个确切的效果而被选择的。我们现在还是可以追求那些因为服务于全面适合度才被植入的那些目标，即使它们现在与全面适合度相冲突——为个体或（至少有

时候)为群体所追求。我们可以使用想象力来构建新的可能性,无论是目标、理论还是冒险的计划,它们都不是根源于以前特定的进化功能。即使想象力本身,即想出新的可能性的能力,是具有进化功能的,我们现在还是能把它用于我们选择的任何目的。

理性人的信念和行动是由一个能可靠地获得某种目标的过程所产生(和维持)的,在那里,理由在指导这种过程中起了某种恰当的作用。在原初什么算作一个理由可以具有进化选择的基础,但是理由提供的指导并不是机械的或盲目的。我们去认知理由,权衡它们,考虑对它们的反驳和它们受削弱的种种方式,然后我们相应地行事和持有信念。

对理由和推理的自我意识为控制和发展增加了另一个维度。哲学通过使得推理本身成为主题而成为第一学科,它使得这种自我意识超越了普通人通常所做的反思。(黑格尔和费希特后来使得自我意识成为他们的主题)。目的和原则得到构建、批判、重构和进一步的发展,并且彼此系统地相关。(此后其他人也承担了哲学家的这个任务,在理论统计学、决策理论和认知科学上产生了海量的文献。) 178

这种发达的理论原则体系——在部分程度上是通过使用进化植入的那些原则所得到的,但不限于此——可以用来指导我们的思想与行为。这个过程也变得自觉了。人们明确发展的这组合理性原则能够应用于那些原则本身——有些应用于其他的,有些应用于自身——这又导致了新的修正和发展。这个弹道能够让我们远离进化的起源。①

① 但存在着这样一种可能性,当合理性用来修正它自身时,合理性的初始缺陷可能不是得到纠正而是被放大了:这个缺陷本身在应用时引起了更大的缺陷。这点可以小心地避免,包括使用这样的外部检验,它的裁定并不自动地确保为肯定的,即使它是从一个经受检验的(公认的)合理原则的修正体系而产生的。

合理性是作为适应一个稳定事实的背景而进化的,它得到选择与该事实相配套起作用。一个这样的事实就是其他具有类似进化合理性的造物的出现。笛卡儿在他的研究中描述了个人单独地决定,他的哪个信念不是错误的或者不是精心欺骗的产物。他的《沉思录》提出了一个程序,这是他的读者在他或她自己的单独研究中也要遵循的。然而,不存在任何理由来认为,进化会这样塑造合理性来符合笛卡儿式个人主义。如果合理性是与他人的并行合理性相伴进化的,那么每个人的合理性都具有一个契合于它的特征,它配合其他人的类似合理性起作用。① 因此,我们不要期待合理性打算证明其他人是理性的或者有能力这样做;为了继续其他事业,这是它所接受且与之共事的因素。

我们的合理性以何种方式来使用他人的合理性呢?我们倾向于从他人那里学习语言,学习年长者表明和告诉我们的各种事实。我们倾向于接受他们所说的,并且接受他们对我们所说东西的纠正,至少直到我们积累了足够的语言和信息,从而能够有根据地怀疑和提出问题时为止。"但是为什么'相信另一个人告诉你的东西'竟然是合理的呢?"我们建立的合理性不是来回答这样的问题,相反它是立基于这种信任,且基于它而得到进一步的发展。如果它是信任的话——它更可能不假思索地把他人(我们首先从之学习的那个群体中的)教导我们看作是理所当然的。

一旦我们在语言和事实性信念上获得了某种基础,我们就

① 参见 R. Boyd and P. J. Richerson, *Culture and the Evolutionary Process* (Chicago: Univ. of Chicago Press, 1985); John Tooby and Leda Cosmides, "Evolutionary Psychology and the Generation of Culture", *Ethology and Sociobiology* 10 (1989): 29-49; Allan Gibbard, *Wise Choices, Apt Feelings*。

可以用之来质疑和修正其他人的信念。合理性的联合进化并不迫使任何人在思想上顺从。如果存在一个要相信其他人的前设,那么这也是能够被克服的。他们是有能力(in a position)去合理地相信一个特定的陈述,还是要从一个有这种能力的人那里学习它呢?个人是一个群体(一个恰当的参照组)的成员,这个群体对某个主题的信念有着种种统计,可以削弱我们的信念前设吗?有任何特殊的理由认为,这个人有动机误导你或者他对准确性没有给予恰当的关注吗?你想出了这个人没有想出的,且与要评估的信念相关的种种可能性吗?在这种人际领域中,我们也有空间来构建、讨论和发展合理信念的进一步原则。[①] 179

语言是合理性的展示与媒介,很多论者都强调了它的社会本质:维特根斯坦论及了判断一致的作用;奎因把语言作为一件社会艺术品加以谈论;普特南描画了语言劳动的分工,那里我们对某些术语的参考是由我们依赖专家知识的方式所决定的。[②] 因为我们的语言能力是与其他人的语言能力相配合而进化的,所有人都是出生在一个有成年说话者的环境之中——我们可以把语言起源的猜测放在一边——如果语言现象和意义竟然是独立于这种社会环境的,那就是件怪事了。

在一个与其他人并存的环境里,进化使得特质的专门化是

---

① 从其他人那里学习时,我们看来设定了他们是理性的——足够理性,可让我们来理解他们正在负责的事情。对于我们已经批判性地讨论过的翻译中的厚道原则而言,存在一种进化基础吗?然而,这样一种原则不一定是如此普遍,以致适用每一个人;只要假定在自己群体里的合理性就够了。

② 参见 Ludwig Wittgenstein, *Philosophical Investigations* (Oxford: Basil Blackwell, 1953); Quine, *Word and Object*; Hilary Putnam, "The Meaning of 'Meaning'", 载于他的 *Mind, Language and Reality: Philosophical Papers*, vol. 2 (Cambridge: Cambridge Univ. Press, 1973), pp. 215-272。

可能的,这些特质被设计得与其他特质一起很好地起作用。如果只有一个人具有亚当·斯密[1]所谈到的那种天生的“物物交易、互换和交换倾向”的话,这种倾向就是没什么用处的。交换倾向需要其他同伴有类似的倾向。正如社会在劳动和技巧专门化上有分工一样,也许一个群体里的生物特征也有分工,有的人更为好斗和尚武,有的人更为灵敏,有的人更为精明,有的人则更为强壮。这是因为人类在生物性上不可能具有所有这些特质,因为与具有互补特质的人生活在一起会对许多人或所有人有益吗?(我宁愿发现这种多样性基础是在个体的选择上,而不是在群体的选择上)。

这看来是讲得通的,但另一种想法更让人不安。合理性本身可能是一组(或由此构成)特质,是否有这种可能性呢?这些特质在群体内具备自然的多样性。在更新世(pleistocene)的狩猎社会里,是否正如进化选择偏爱更强壮和更灵敏的人那样,进化也选择偏爱那些信念和计算方面更理性的人呢?是不是几乎所有人都得益于这些混合特性才使其得以可能的那些合作与交换活动呢?也许合理性显示出的各种差别完全出自于非生物性180 的原因?无论起因是什么,不管原因是什么,人们会情不自禁地认为,更理性者会毫不犹豫地主张他们自己具有更高贵的地位,即便只是因为他们更擅长于言语,由此能够更好地阐述及捍卫他们的这个主张。对于极度理性者的这种自我溢美之词,毫无疑问,其他人是置之不理的。

但是,事实上则是合理性已经重塑了世界。这是韦伯作品中的伟大论点:经济和货币计算、官僚制理性、普遍规则和程序

---

① Adam Smith, *The Wealth of Nations*, bk. 1, ch. 2.

已经代替了基于个人纽带的行动，市场关系扩展到了新领域。[1]与明确地利用了且依赖于合理性的相关制度变化一起，合理性的收获颇丰且使得合理性有能力进一步扩大其领域。

然而，这也使得这个世界以各种各样的方式敌视较低程度的合理性。其传统对于韦伯式合理性不大感冒的文化就过得不那么如意了。在西方社会里，曾经服务于狩猎社会的那种特质分工上的平衡已经被打破了。合理性首先通过给其他特质带来好处而能够扩展其影响力，但这些特质也变得越来越依赖合理性，合理性也变得越来越强大，受到的约束也越来越少。合理性现在是在重塑世界来适应它自己，改变的不仅仅是它自己的环境，而且也改变了其他特质所处的环境，扩展了只有它才能充分繁荣的环境。在那个环境里，合理性的边际产品增加了，而其他特质的则减少了；一度是处于合作地位的其他特质现在处于低级地位了。这对于合理性的激情(compassion)和想象力与原创性是个挑战：它有能力设计出这样的系统，那里各种特质能够和谐共处、一起繁荣(如果它们愿意的话也有机会去发展它们的合理性)吗？它又愿意这样做吗？

柏拉图谈到了理解永恒形式，亚里士多德论及了对第一原则的智力直觉和心灵的认知本质，笛卡儿谈到了沐浴在自然理性之光下的清晰且独特的理念和真理。我们进化论说的读者很可能会纳闷合理性的尊严究竟变成了什么东西。当然，合理性的贬低性论说不是个新鲜事；休谟跟康德也分配给合理性一个更有限的功能。康德剥夺了“思辨理性假装具有的那种超验洞察力”，他说，“为了给信仰以空间。”然而，知道了合理性的起源和初始功能，这并不会夺掉合理性所有的高贵性。(我们最好还

[1] Max Weber, *Economy and Society* (New York: Bedminster, 1968).

是记住,高贵性,尽管它们频繁地宣称,但实际也没有特殊的起源。)考虑一下好奇心。即使某种程度的好奇心得到选择是因为181 它在产生具有实践用处的新真理上的作用,但一旦这样一种能力存在了,它就能转而探查宇宙的起源、无穷的本质,地球生命的起源和发展,还有人性的范围和局限,所有这些都是为了满足智力好奇心本身和为了由此带来的知识,不再有进一步的驱动性目的。如果理性不是一个对独立实体的不可错的认知者,那么这也许使得它的胜利更让人震惊和印象深刻。无论美学辨别的实践起源可能是什么,它一直被用来产生伟大的艺术作品。当最崇高的人类创造似乎是派生于低端的起源和功能时,需要修正的不是我们对于那些创造的敬意,而是我们的高贵性概念。能促使我们得到最崇高的人类成就的起点又能有多低,这样一种潜能和力量的起源又能有多么卑贱呢?

对于我们的行动、情感以及这个世界,合理性让我们有了更多的知识和更大的控制。尽管我们的合理性在起初是个进化的性质——合理性的本质(nature)把自然(Nature)也包含在合理性之中——它使得我们转变我们自己,因此超越了我们作为单纯动物的身份,既是实际地也是象征地。合理性开始塑造和控制它自己的功能。

我们的原则确定了我们的生活代表了什么,我们的目的创造了我们的生活所沐浴于其中的光明,我们的合理性,既有个体的也有合作的,定义和象征了我们与单纯动物性之间的距离。正是这些手段使得我们的生活的意义不止是它们工具地产生的东西。意指的越多,我们的生活生产得就越多。

主 题 索 引

（后面所标页码为原书页码，即本书边码；
注释所指页码为本书页码）

人 名 索 引

(后面所标页码为原书页码,即本书边码;
注释所指页码为本书页码)

译 后 记

本书的翻译缘于陈昉博士，她非常粗糙地翻译了《合理性的本质》，但既没时间也没有兴趣来继续这个任务。她熟知我对诺奇克兴趣颇大，也颇为相信我的学术态度，由此决定把这个粗糙的草译稿交给我，任由我处置。目前的版本就是在她的译文基础上对全书所进行的重译。尽管这本书的篇幅不大，而且诺奇克的行文也非常清楚、简洁，但其翻译难度远远超出了我的预估。原因无他，诺奇克在书中随意地穿梭于各个学科，从经济学到量子力学，旁征博引，以此构建与支持他自己的合理性理论。这不仅增加了理解的难度，更增加了翻译的难度。一是不大知道各个学科约定俗成的定译，得多方查证，另外则是有些术语在不同学科定译还不尽相同，这都增加了处理的难度。

为此，本书的翻译不仅花费了我很多的时间，而且得到了许多人的帮助才达到目前这个样子。这里首先要感谢思勉人文高等研究院所提供的自由空间，这里没有对研究人员在科研上做出硬性规定，这才使得我能够投入充足的时间来完成这个并不被目前的“学术体制”所认可的翻译工作。同时，译者在这里也要感谢好友潘军伍、加尔文学院的 Wykstra 教授，还有天涯社区的英语杂谈中许多不知名网友的热心帮忙，他们均帮助我解决了一些疑难句子的理解，由此避免了不少的错误。此外，李河编辑对我在《世界哲学》上发表的译文所提出的修订意见，还有

加尔文学院的 Wykstra 教授所上的英文写作课，都使得我更为注意，如何从一个读者的角度来思考译文，把译文处理得更为顺畅，更易理解。这里还要感谢我的妻子宋学芳女士，为了克服我译文的“西化”现象，同时帮助检验译文的可读性，她以一个普通读者的视角，硬着头皮阅读了译文的前三章，提出了不少译文表达上的问题。这里还要感谢邓正来先生允许使用他与陈昉博士合译的导论部分。最后，译者还要感谢应奇老师热心帮忙联系出版社，同样感谢译文出版社马胜先生的信任，使得我能够有机会翻译这本书，而正孕育着“希望”的王巧贞女士，也为此书付出了大量的辛苦劳动，相信读者在阅读过程中会验证这一点的，这自然也是译者需要感谢的。尽管我已经想了各种办法来改善译文质量，但由于水平有限，最终错漏之处在所难免，恳请方家批评指正，以期以后对译文加以改正与完善。但无论怎样，一切翻译责任均由译者个人承担。

译　者

于人文楼